Thesaurusgestützter Zugriff zu Umweltberichten in einem netzübergreifenden Hypertextsystem

dem Fachbereich Mathematik, Naturwissenschaften und Informatik
der Fachhochschule Gießen-Friedberg
vorgelegte Diplomarbeit zur Erlangung des Grades
Diplom-Informatikerin (FH)

von
Margit Gaul
geb. am 19.01.1968 in Höchst

Referent : Prof. Dr. Klement
Koreferent : Prof. Dr. Hoffmann

Gießen, Februar 1995

Inhaltsverzeichnis

1. Einleitung

Das Bewußtsein für Umweltfragen nimmt in der Welt einen immer größeren Raum ein. In dieser Welt, die immer bessere und raschere Informationsflüsse schafft, wächst parallel zu diesem Umweltbewußtsein die Realisierung wissenschaftlicher Systeme, die die Nachfrage nach Informationen befriedigen können. Konzepte wie "Hypertext" sind dazu angelegt, einer breiten Öffentlichkeit einen benutzerfreundlichen Zugang u.a. auch für Umweltdaten zu ermöglichen.
Hypertext ist ein Gebiet, das, besonders in Zusammenhang mit World Wide Web, in den letzten Jahren zunehmend an Interesse gewinnt, und das nicht nur in "Computerkreisen". Hypertext erlaubt durch einfache Aktionen das assoziative "Wandern" in Texten, indem beim Anklicken bestimmter Textstellen mit der Maus weitere relevante Informationen zu diesen Textstellen angezeigt werden. Durch World Wide Web können im Internet vorhandene Informationen als Hypertext über eine besonders einfach zu handhabende Benutzeroberfläche zugänglich gemacht werden. Dieser einfache Zugang erscheint auch im Hinblick auf die EWG-Richtlinie vom 07. Juni 1990 über den freien Zugang zu Umweltinformationen besonders wichtig.

Die vorliegende Arbeit beschäftigt sich damit, wie Umweltberichte als Hypertext aufbereitet und im World Wide Web zur Verfügung gestellt werden können. Der Zugriff auf diese Umweltinformationen ist dabei nicht nur über die Texte selbst gegeben, sondern auch über in einem Thesaurus abgelegte Schlagwörter aus dem Umweltbereich.

Der praktische Teil der Arbeit entstand im Rahmen des Umweltinformationssystems Baden-Württemberg (UIS) als ein Teil des GLOBUS-Projektes am Forschungsinstitut für anwendungsorientierte Wissensverarbeitung (FAW) an der Universität Ulm. GLOBUS wurde im Auftrag des Umweltministeriums Baden-Württemberg für die Landesanstalt für Umweltschutz (LfU) entwickelt. Projektpartner waren neben dem FAW das Institut für Photogrammetrie und Fernerkundung (IPF) der Universität Karlsruhe, das Institut für Kernenergetik und Energiesysteme (IKE) der Universität Stuttgart sowie das Forschungszentrum Informatik (FZI) an der Universität Karlsruhe. Inhalt des GLOBUS-Projektes war die »Konzeption und prototypische Realisierung einer aktiven Auskunftskomponente für globale Umwelt-Sachdaten«. Diese Auskunftskomponente, vorrangig für Referenten im Umweltministerium und in der LfU gedacht, soll unterschiedliche Datenbestände erschließen und auf verschiedenen Rechner-Plattformen verfügbar sein.

Der Datenbestand, der im Rahmen dieser Arbeit erschlossen wurde, ist der vom Umweltministerium Baden-Württemberg herausgegebene Bericht "Umweltdaten 91/92"; für den thesaurusgestützten Zugriff auf diesen Umweltbericht wurde der Umwelt-Thesaurus des Umweltbundesamtes integriert.

Gegenüber der gedruckten Form des Umweltberichtes bietet die Hypertext-Version ein wesentlich breiter gefächertes Spektrum an Informationsmöglichkeiten. Ungezieltes Stöbern in den Texten wird dabei genauso unterstützt wie gezieltes Informieren über konkrete Umweltbelange.
Der Zugang zum Umweltbericht ist zum einen über Inhaltsverzeichnisse, zum anderen aber auch über die Begriffe aus dem Thesaurus möglich, so daß gezielt Dokumente zu bestimmten Themen gesucht werden können. Dazu müssen keine komplexen Suchanfragen formuliert, sondern nur der entsprechende Begriff im Thesaurus angeklickt werden. Zusammenhängende Informationen können auch vom Umweltbericht ausgehend lokalisiert werden, indem die jedem Unterkapitel des Umweltberichtes zugeordneten Schlüsselbegriffe angeklickt werden. Für jeden dieser Schlüsselbegriffe existiert eine Übersichtsseite, die alle zu diesem Begriff in Beziehung stehenden Unterkapitel des Umweltberichtes anzeigt. Zusätzlich kann hier zu anderen Begriffen verzweigt werden, die zu diesem Schlüsselbegriff in einem Zusammenhang stehen (z.B. Unterbegriffe oder verwandte Begriffe). So ist es praktisch für jedermann möglich, auch ohne Kenntnis einer speziellen Abfragesprache Informationen zu einem bestimmten Thema zu finden.
Für speziellere Suchanfragen, z.B. die Verknüpfung von mehreren Suchbegriffen durch eine "Und"-Beziehung, besteht zusätzlich die Möglichkeit zu einer Thesaurus-unabhängigen Volltextrecherche.

Zur Beschreibung der Arbeit werden im methodischen Teil zunächst Grundlagen von Hypertext und des Information Retrieval (speziell der Indexierung) erläutert. Anschließend wird auf die Verbindung dieser beiden Gebiete eingegangen und ein konkretes Werkzeug, World Wide Web, beschrieben.
Der praktische Teil beginnt mit einer Spezifizierung der Anforderungen an das zu entwickelnde System. Bevor auf dessen Entwurf und Realisierung eingegangen wird, wird zum besseren Verständnis eine Beschreibung des Systems aus Benutzer- bzw. aus Autorensicht vorangestellt. Daran schließt sich eine Darstellung der verwendeten Eingabedaten an. Im Systementwurf wird zunächst das Gesamtsystem mit seinen Teilkomponenten im Überblick beschrieben, danach wird die konkrete Realisierung dieser Teilkomponenten vorgestellt.
Die abschließende Diskussion geht auf mögliche Verbesserungen und Erweiterungen des entwickelten Systems ein.

2. Methodische Grundlagen

2.1 Hypertext

2.1.1 Hypertext und Hypertextsysteme

Zu Hypertext bzw. Hypermedia gibt es in der Literatur viele unterschiedliche Definitionen, wobei die Begriffe Hypertext und Hypermedia meist nicht klar getrennt sind. Gewöhnlich wird zwar zunächst Hypertext von Hypermedia abgegrenzt bzw. Hypermedia als Erweiterung von Hypertext auf andere Medien als Text angesehen, aber dann werden die Begriffe üblicherweise synonym benutzt. Andere Definitionen schließen bei der Beschreibung von Hypertext Hypermedia mit ein, was auch für diese Arbeit gelten soll.

So unterschiedlich die Definitionen auch sind, ist doch der zentrale Punkt aller Definitionen die nicht-lineare Verknüpfung von Informationseinheiten.
Durch diese Nicht-Linearität wird "Hypertext" von "Text" abgegrenzt. Zwar weist auch jeder Text Nicht-Linearität (in Form von Inhaltsverzeichnissen, Fußnoten, Querverweisen, etc.) und jeder Hypertext Linearität auf, aber "[...] das Grundprinzip von Text [ist] Linearität und das von Hypertext Nicht-Linearität" (KUHLEN 1991).
Eine häufige Darstellung von Hypertext ist die eines Netzwerkes, in dem die einzelnen Knoten (die Informationseinheiten) Text, Graphik oder multimediales Material repräsentieren und die Kanten die Beziehungen zwischen diesen Objekten darstellen. Die Verknüpfungen zwischen diesen Objekten werden auch als Verweise, Hypertext-Verweise, Hypertext-Links, Hyperlinks oder kurz Links bezeichnet. Unterscheiden kann man diese Verknüpfungen nach KUHLEN (1991) danach, ob sie

- Ausgangs- und Zielpunkte innerhalb von Einheiten verbinden (intrahypertextuelle Verknüpfungen)

- Ausgangs- und Zielpunkte zwischen verschiedenen Einheiten verbinden (interhypertextuelle Verknüpfungen)

- vom Ausgangspunkt (in der Hypertextbasis) auf Zielpunkte in externen Objekten (z.B. eine andere Hypertextbasis oder auch ein ganz anderes Informationssystem) verweisen (extrahypertextuelle Verknüpfungen).

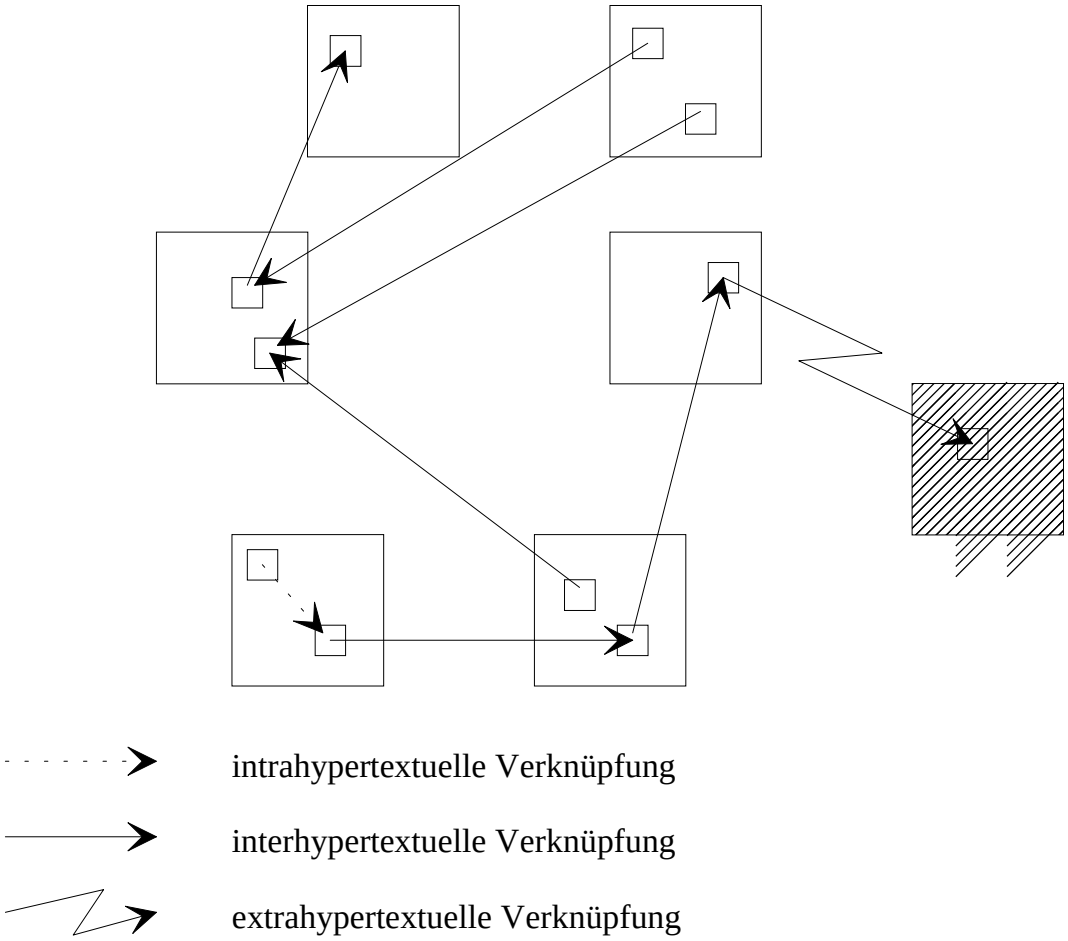


Abb. 1 Hypertext als Netzwerk

SCHNUPP (1992) definiert Hypertext in Anlehnung an Nelson als "[...] eine Menge von textuellem oder bildlichem Material, das so komplex vernetzt ist, daß es nicht auf einfache Weise auf Papier dargestellt oder vermittelt werden kann".
Entsprechend dieser Definition versteht man dann unter einem Hypertextsystem "[...] ein (Software-)System, das die Erstellung und Nutzung derartiger vernetzter Textinformationen unterstützt" (SCHNUPP 1992).

Ein Hypertextsystem besteht aus einer Hypertextbasis (auch bezeichnet als Hypertext oder als Hyperdokument), einem Hypertextmanagementsystem sowie einer Autoren- und einer Navigationskomponente.

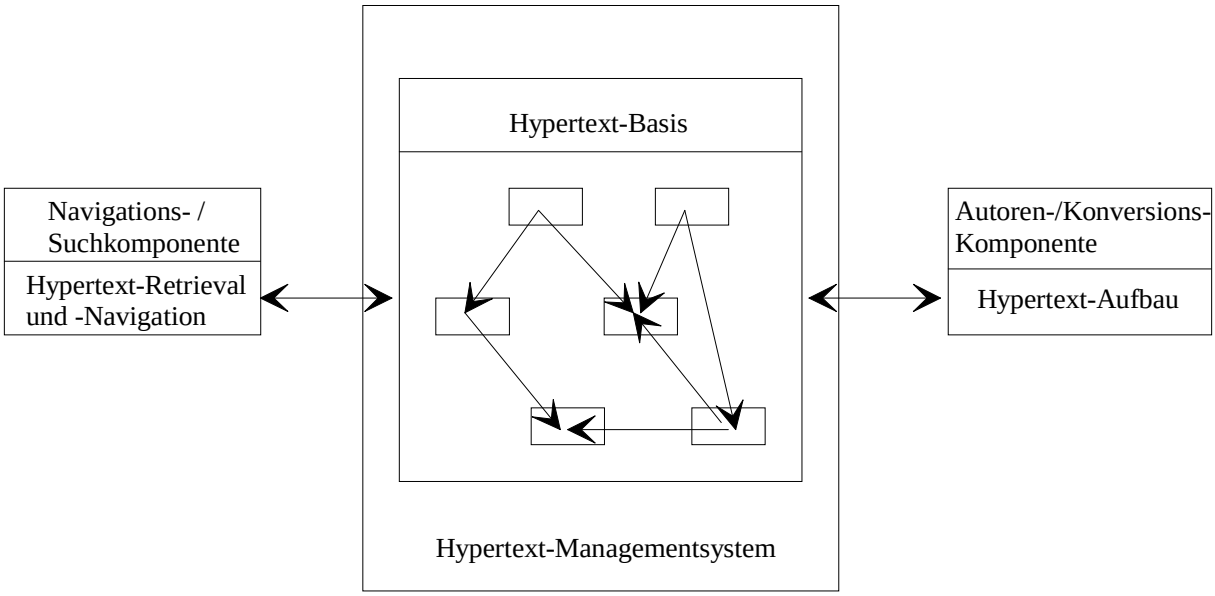


Abb. 2 Aufbau eines Hypertextsystems (nach KUHLEN 1991)

Für die Navigationskomponete (oder das "Lesesystem") wird auch die Bezeichnung "Browser" verwendet. Da sich für "Browser" bzw. "Browsing" noch keine entsprechende deutsche Bezeichnung durchgesetzt hat (verwendet wird manchmal "Stöbern" oder "Durchblättern"), sollen hier die englischen Begriffe verwendet werden. Eine sehr treffende Definition für "Browsing" findet sich bei RADA (1991), der es (nach COVE und WALSH 1988) bezeichnet als "the art of not knowing what you want until you find it".

Die Autoren-/Konversionskomponente dient dazu, aus einem Text einen Hypertext aufzubauen, d.h. entsprechende Einheiten mit Verknüpfungen zu erstellen. Dies kann manuell oder über entsprechende Programme geschehen.

Das Hypertext-Managementsystem ist zuständig für die Verwaltung der Hypertextbasis (z.B. über Datenbanken oder Dateiverwaltungssysteme).

2.1.2 Geschichte von Hypertext

Im folgenden soll kurz die Geschichte von Hypertext (d.h. eigentlich eine Geschichte der Hypertextsysteme) in Anlehnung an NIELSEN (1990) dargestellt werden.

Der Begriff "Hypertext" wurde erstmals 1965 von Ted Nelson verwendet. Als Idee existiert Hypertext jedoch schon wesentlich länger. Das erste Hypertextsystem wurde von Vannevar Bush (1890-1974) Anfang der 30er Jahre konzipiert und 1945 in seinem Aufsatz "As We May Think" (BUSH 1945), dem die Anfänge der Hypertext-Idee zugeschrieben werden, beschrieben. Dieses System MEMEX ("memory extender") sollte auf Mikrofilmbasis

realisiert werden, wurde jedoch nie implementiert. Durch eine Kamera sollte alles Interessante aufgenommen und sofort im MEMEX verfügbar gemacht werden. Über eine Art Bildschirm sollte dann der Zugang zu diesen auf Mikrofilm gespeicherten Informationen möglich sein (s.a. KUHLEN 1991).
Gemäß NIELSEN (1990) kommt der Grundgedanke von Hypertext ("Hypertext, in other words !") besonders im folgenden Zitat von BUSH (1945) zum Ausdruck:

> All this is conventional, except for the projection forward of present-day mechanisms and gadgetry. It affords an immediate step, however, to associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of the memex. The process of tying two items together is the important thing.

Nach diesem Aufsatz tat sich auf dem Gebiet "Hypertext" längere Zeit nicht viel. 1962 wurde am Stanford Research Institute mit der Arbeit am Projekt "Augment", das in den Bereichen Büroautomation und Textverarbeitung angesiedelt war, begonnen. Teil dieses Projektes war NLS (oN-Line System), das zwar nicht als Hypertextsystem konzipiert war, aber dennoch verschiedene Hypertext-Merkmale trug. Während des Augment-Projektes wurden die Aufzeichnungen der Mitarbeiter in einer "Journal"-Komponente gespeichert, die es ihnen erlaubte, Verweise zu anderen Arbeiten in ihren eigenen Texten einzubinden. Auf diese Weise wuchs das Journal auf über 100.000 Einträge. Interessant zu erwähnen ist in diesem Zusammenhang noch, daß zur Unterstützung des Augment-Systems die Maus erfunden wurde (s. RADA 1991).

Nelson führte den Begriff "Hypertext", wie bereits oben erwähnt, 1965 ein, als er mit den Arbeiten am Hypertext-Projekt "Xanadu" begann.
Zwar gibt es seit 1990 ein Produkt "Xanadu", aber die eigentliche Vision von Nelson, nämlich die eines universellen, weltweit realisierten Hypertextes, der alles bisher Geschriebene verknüpft, wurde bisher noch nicht verwirklicht. Dennoch gibt es heute ein anderes System, das dieser Zielvorstellung, "die physikalische Präsenz relevanter Information am Arbeitsplatz zugunsten der logischen Verknüpfung zu beliebig entfernten Einheiten jeder medialen Art aufzuheben" (KUHLEN 1991), sehr nahe kommt, nämlich "World Wide Web" (s. 2.5). Hierzu bemerken auch DECEMBER und RANDALL (1994): "[Nelson] envisioned a system of information access not much different from what we have today, and there's little doubt that he must cruise the World Wide Web with a knowing smirk and an attitude of 'I told you so'."

Das erste funktionierende Hypertextsystem war das 1967 an der Brown University entwickelte "Hypertext Editing System", das nach dem Abschluß des Forschungsprojektes an das Houston Manned Spacecraft Center verkauft wurde. Dort wurde es genutzt, um Dokumentationen für die Apollo-Missionen zu erstellen.

Als Fortsetzung des "Hypertext Editing System" folgte 1968 das ebenfalls an der Brown University entwickelte FRESS (File Retrieval and Editing System), wie auch schon sein Vorgänger auf einem IBM-Rechner implementiert.
Beide Systeme wiesen die grundlegende Hypertext-Funktionalität (d.h. Verweise und Springen zu anderen Dokumenten) auf, jedoch war die Benutzerschnittstelle größtenteils textorientiert und verlangte eine indirekte Spezifizierung der Sprünge.

Weitere wichtige Stufen in der Entwicklung von Hypertext waren die "Aspen Movie Map" (1978), das wahrscheinlich erste Hypermediasystem, und "Filevision" (1984), ebenfalls ein Hypermediasystem.
Im Gegensatz zu den frühen Hypertextsystemen, die meist nur für den "Eigengebrauch" entwickelt wurden, war der "Symbolics Document Examiner" (entwickelt ab 1982) als Produkt für die Benutzer der Symbolics Workstations konzipiert, und zwar als Schnittstelle zur Online-Dokumentation der Workstations. Die Hypertextversion dieser auch in gedruckter Form vorliegenden Dokumentation (ca. 8.000 Seiten) umfaßte 10.000 Knoten mit 23.000 Hyperlinks. Der "Document Examiner" wird als das erste Hypertextsystem in "real world use" betrachtet, aber aufgrund der hohen Kosten der Symbolics Workstations war es nicht sehr weit verbreitet.
Anders dagegen das System "Guide" (1986), das in der ersten Version auf Macintosh-Rechnern und später auch auf dem IBM PC lief.
Populär wurde Hypertext dann 1987 mit Einführung des Programmes "HyperCard" durch Apple; das Programm wurde kostenlos mit jedem Macintosh ausgeliefert. In diese Zeit fällt dann auch die erste ACM-Konferenz zu Hypertext, die seither regelmäßig fortgesetzt wird.

Um die gesamte Entwicklung von Hypertext mit NIELSEN (1990) zusammenzufassen:

> In conclusion, we can say that hypertext was conceived in 1945, born in the 1960s, and slowly nurtured in the 1970s, and finally entered the real world in the 1980s with an especially rapid growth after 1985, culminating in a fully established field during 1989.

2.2 Aufbau von Hypertexten aus Texten

In Zukunft werden zwar vermehrt Hypertexte ohne den Umweg über Text erstellt werden, aber momentan stellt sich meist das Problem, bestehende Datenbestände in Hypertextsysteme zu integrieren. Dies kann, je nachdem in welchem Format diese Datenbestände vorliegen, mehr oder weniger schwierig sein, wird aber immer eine Vielzahl von Fragen und Möglichkeiten einschließen.

Grundsätzlich sollte beim Aufbau von Hypertexten, unabhängig von der Art der Konversion, eine Anpassung an die Möglichkeiten und Vorteile von Hypertext angestrebt werden, anstatt Textstrukturen zu imitieren (s. KUHLEN 1991).
In den nächsten Abschnitten sollen verschiedene Konversionsformen sowie die Vorgehensweise bei der automatischen Konversion, insbesondere die Generierung von Hyperlinks, dargestellt werden.

2.2.1 Konversionsformen

Die erste grundsätzliche Unterscheidung beim Aufbau von Hypertexten aus Texten kann man zwischen manueller und automatischer Konversion treffen. Nach RINER (1991) ist die manuelle Konversion machbar für Prototypen, aber sobald die Texte einige Seiten übersteigen, ist sie nicht mehr denkbar. In der Realität wird man wahrscheinlich eher eine Mischform von manueller und automatischer Konversion antreffen, d.h. es wird bei den meisten automatischen Konversionsverfahren eine manuelle Nachbearbeitung der erstellten Dokumente und Hyperlinks nötig sein.

KUHLEN (1991) unterscheidet folgende Konversionsformen:

- Einfache Übertragung
  Bei dieser Konversionsform wird der Text 1:1 in eine Datei eines Hypertextsystems übertragen, also z.B. eine Datei aus einem Textverarbeitungsprogramm in eine HTML-Datei konvertiert. Zu einem Hypertext wird der Text dann durch den Aufbau nicht-linearer Strukturen, die entstehen, wenn verschiedene Elemente oder Passagen im Text miteinander verknüpft werden. Durch diese Art der Konversion findet keine Segmentierung des Textes statt, vielmehr wird ein hypertextgerechtes Browsing durch den gesamten Text ermöglicht.

- Segmentierung und Relationierung über formale Texteigenschaften
  Üblich ist, den Ausgangstext nicht wie bei der einfachen Übertragung als eine Einheit zu belassen, sondern ihn in hypertextgerechte Einheiten zu zerlegen (Segmentierung). Dazu werden formale Strukturen im Text, wie z.B. Kapitel oder Unterkapitel, ausgenutzt. Die Erzeugung nicht-linearer Strukturen (in diesem Fall die Verknüpfung der segmentierten Einheiten) kann ebenfalls über formale Texteigenschaften erfolgen.

- Segmentierung und Relationierung nach Kohärenzkriterien
  Durch diese Art der Konversion wird der Ausgangstext reorganisiert, d.h. es werden evtl. neue Hypertext-Einheiten erstellt, die in dieser Form im Ausgangstext noch nicht

vorhanden waren. Voraussetzung dafür ist eine inhaltliche Analyse unter Ausnutzung semantischer und argumentativer Eigenschaften von Texten.

- Intertextuelle Konversion
  Von einer intertextuellen Konversion spricht man, wenn eine Hypertextbasis aus einer Vielzahl von Texten aufgebaut wird. Im Prinzip können hier die gleichen Techniken wie bei der Übernahme eines einzelnen Textes verwendet werden. Vorher sollte jedoch geprüft werden, ob durch die Kombination verschiedener Texte ein informationeller Mehrwert gegenüber mehreren Einzelversionen für den Benutzer entsteht.

- Einbindung von textuellen Strukturmitteln
  Neben dem eigentlichen Text müssen auch die textuellen Strukturteile (z.B. Inhaltsverzeichnisse, Register) übernommen und angepaßt werden.

2.2.2 Schritte bei der automatischen Konversion

Die einzelnen Schritte bei der Konversion beschreibt z.B. RINER (1991). Da sich seine Vorgehensweise teilweise auf ein bestimmtes Werkzeug (HyperTRANS von Texas Instruments) bezieht, sollen hier nur die Punkte herausgegriffen werden, die von grundsätzlichem Interesse erscheinen.

- Auswahl geeigneter Texte
  Hier stellt sich die Frage, welche Texte am ehesten von einer Konversion in Hypertext profitieren. Geeignet sind z.B. Texte, die viele Querverweise enthalten und/oder eine komplexe Struktur aufweisen.
  Diese Problematik wird auch bei KUHLEN (1991) aufgegriffen. Er nennt als Beispiele für zur Konversion geeignete Texte u.a. solche, deren Inhalt sich leicht in Blöcke aufteilen läßt, die viele strukturelle Metainformationen (Register, Abkürzungsauflösungen,...) enthalten sowie Texte mit weitgehend abgesicherten Wissensstrukturen.

- Behandlung von Graphiken
  Gesichtspunkte, die bei der Konversion und Einbindung von Graphiken berücksichtigt werden sollten, sind z.B. die Auflösung des gedruckten Bildes und der Bildschirm-Version, die Größe der Graphiken (können sie z.B. auf die Größe einer Bildschirmseite verkleinert werden) und die Frage, ob Elemente in der Graphik zu anderen Textstellen/Graphiken verweisen sollten.

- Umsetzung der Texte in maschinenlesbare (vom Konvertierungsprogramm lesbare) Form
  Grundsätzlich kommen hier zwei Möglichkeiten in Betracht, nämlich zum einen das Einscannen des Originaldokumentes und zum anderen die Verwendung bereits existierender Dateien. Scannen ist notwendig, falls das Dokument gar nicht als Datei vorliegt; selbst wenn Dateien vorliegen, ist es oft so, daß die Papierfassung die aktuellste Version ist.
  Bei der Verwendung von Dateien kann auch eine Vorverarbeitung nötig sein, falls sie in einem Format vorliegen, daß nicht direkt für eine Konvertierung in Hypertext geeignet ist. Dann müssen die Dateien in ein Standard-Format gebracht werden, das als Eingabe für die Konvertierung akzeptiert wird.

- Starten des Konversionsprozesses
  Nachdem die Eingabedokumente in für die Konvertierungssoftware passendem Format vorliegen, kann der Konversionsprozeß gestartet werden. Dieser muß folgende Aufgaben erledigen:
  - Aufteilen des Eingabedokumentes in Knoten
  - Aufbau hierarchischer Links vom Inhaltsverzeichnis zu einzelnen Kapiteln, von dort zu Unterkapiteln usw.; evtl. muß ein Inhaltsverzeichnis erst noch automatisch erstellt werden
  - Aufbau nicht-hierarchischer Links zwischen den Knoten, die auf Verweisen zu Kapiteln, Abbildungen, Fußnoten usw. basieren
  - Aufbau externer Verweise zu anderen Dokumenten.

- Fehler in Eingabedokumenten
  Eine wichtige Frage bei der automatischen Konversion ist die Behandlung von Fehlern in den Eingabedokumenten. Diese können einfache Schreibfehler, aber auch inkonsistente Strukturen sein; je nach Fehlerart kann die Behebung durch die Konvertierungssoftware geschehen oder muß durch den Autor des Textes erfolgen.

- Komplettierung der Konversion
  Nach der Fehlerkorrektur muß die erzeugte Hypertextbasis überprüft werden. Dies betrifft besonders die Hyperlinks, bei denen geprüft werden muß, ob alle einen Ursprung und ein Ziel haben, bzw. ob sie inhaltlich richtig sind.

2.2.3 Automatische Generierung von Hyperlinks

Ein sehr wichtiger Aspekt (vielleicht der wichtigste) bei der automatischen Konversion ist die automatische (bzw. computerunterstützte) Generierung der Hyperlinks. Auch in der Literatur finden sich viele Beiträge zu diesem Problem. Als Beispiele seien hier BERNSTEIN (1990), KNORZ (1992), REARICK (1991) und WILSON (1990) genannt.

Ein Ansatz bei der automatischen Generierung von Hyperlinks besteht darin, formale Texteigenschaften auszunutzen. Beispielsweise können so aus gegebenen Inhaltsverzeichnissen Hyperlinks zu den einzelnen Kapiteln generiert oder aus Überschriften Inhaltsverzeichnisse und dann Hyperlinks erstellt werden. Gut ausnutzen lassen sich auch im Text enthaltene Hinweise wie z.B. "siehe auch", "siehe Abb.", Fußnoten oder Verweise auf Literatur. Diese können direkt in Hyperlinks umgesetzt werden.
Eine weitere Methode, die gegenüber den unten beschriebenen ausgeklügelten Analysetechniken mehr "straigt forward" ist, ist die Verwendung von Strukturelementen des zugrundeliegenden Textsystems (praktisch jeder Text wird heutzutage mit einem Textsystem erstellt). Formatvorlagen (in Word) oder Kommandos (in TeX) ermöglichen es, Überschriften, Fußnoten, Bildunterschriften, Literaturverweise, Querverweise usw. automatisch zu erkennen, weil der Autor diese gekennzeichnet hat. Es gibt Beispiele für derartige Konverter.

Diese Beispiele sollen nur einen Eindruck vermitteln, auf welche Methoden man bei der automatischen Generierung von Hyperlinks zurückgreifen kann. Eine systematische Einteilung findet sich bei REARICK (1991), der zwischen lexikalischer, statistischer, syntaktischer und semantischer Analyse unterscheidet. Im folgenden sollen diese Ansätze kurz zusammengefaßt werden:

- Lexikalische Analyse
  In einer einfachen Form kann die lexikalische Analyse auf Zeichenkettenvergleichen beruhen. Dadurch können z.B. Hyperlinks von jedem Vorkommen eines Wortes zu dessen Definition erzeugt werden. Nachteil ist, daß auf diese Weise viele redundante Links erzeugt werden, die allerdings verringert werden könnten, indem der Hyperlink nur beim ersten Vorkommen eines Wortes in einem Absatz angelegt wird. Schwierigkeiten ergeben sich bei der lexikalischen Analyse auch durch die verschiedenen Wortformen, die erst durch einen Algorithmus auf eine gemeinsame Stammform reduziert werden müssen. Weitere Beschränkungen ergeben sich durch Synonymie (verschiedene Wörter mit gleicher Bedeutung) und Polysemie (verschiedene Bedeutungen für ein Wort).

- Statistische Analyse
  Aufgrund der Ergebnisse von statistischen Analysen der Worthäufigkeit und -verteilung innerhalb eines Textes können Hypertext-Indizes und interne Querverweise automatisch generiert werden. Ein Hypertext-Index kann dabei mit den Methoden der automatischen Indexierung aufgebaut werden, allerdings mit Einschränkungen, da das Ziel bei Hypertext-Indizes ein anderes als bei Indizes zum Text-Retrieval ist und damit andere Voraussetzungen gegeben sein können. Aus diesen Gründen ist eine automatische Indexierung, die lediglich auf Worthäufigkeiten beruht, nicht unbedingt sinnvoll; Verbesserungen können durch Stoppwort- und Deskriptorlisten erzielt werden.

- Syntaktische Analyse
  Durch eine Syntaxanalyse können z.B. verschiedene Bedeutungen eines Wortes unterschieden werden (je nach der Stellung im Satz kann ein Wort z.B. einmal als Substantiv und einmal als Adjektiv verwendet werden).
  Außerdem kann die hierarchische Struktur von Kapiteln und Abschnitten syntaktisch beschrieben werden. Durch eine Analyse können dann z.B. automatisch Inhaltsverzeichnisse und die entsprechenden Hyperlinks generiert werden.

- Semantische Analyse
  Im Gegensatz zu den drei zuvor beschriebenen Ansätzen, die lediglich Zeichenketten analysieren, sollen durch die semantische Analyse Hyperlinks, die auf der Bedeutung dieser Zeichenketten basieren, generiert werden. Dieser Ansatz ist umstritten, da Programmen zur semantischen Analyse wiederum einer oder mehrere der nicht-semantischen Ansätze zugrunde liegen.

Ein für die weitere Arbeit wichtiger Ansatz soll noch einmal herausgegriffen werden, und zwar die Möglichkeit, Hyperlinks mit Hilfe von automatischer Indexierung zu erstellen.
Dies wird auch bei KNORZ (1992) angesprochen ("Es liegt auf der Hand, daß Techniken des automatischen Indexierens die Basis einer automatischen Link-Generierung bilden könnten"). Außerdem wird dort auf die Möglichkeit einer Funktionalität hingewiesen, die es erlaubt, von einer Textstelle zu weiteren Textstellen zu verzweigen, in denen dieselbe Wortform vorkommt. Dies alles führt zum Problem des Indexierens, das Gegenstand des nächsten Abschnittes ist.

2.3 Indexierung

Unter "Indexierung" versteht man "[...] die inhaltliche Kennzeichnung eines Dokuments durch Zuteilung von Deskriptoren für die Zwecke des Dokumenten-Retrieval. Indexierung bezeichnet sowohl den entsprechenden Vorgang, d.h. die Zuteilung der Deskriptoren, als auch das Ergebnis, d.h. die Gesamtheit der einem Dokument zugeordneten Deskriptoren [...]" (LUSTIG und ZIMMERMANN 1991).

Das Indexieren kann, je nach Vorgehensweise, nach verschiedenen Kriterien unterschieden werden. Im folgenden sollen einige der gebräuchlichsten Unterteilungen vorgestellt werden.
Eine erste wichtige (in der Literatur zu Information Retrieval meist zu findende) Unterscheidung kann man zwischen manuellem (nichtautomatischem, intellektuellem) und automatischem Indexieren treffen. Allerdings wird hier oft unterschiedlich abgegrenzt. Während z.B. SALTON/MCGILL (1987) schon von automatischer Indexierung sprechen, wenn die Indexierung mit dem Computer durchgeführt wird, verstehen LUSTIG/ZIMMERMANN (1991) darunter "[...] die Automatisierung des Indexierens einschließlich der Entwicklung der dafür erforderlichen Wörterbücher". Für die weitere Arbeit soll darunter die automatische Extrahierung von Deskriptoren aus den Texten verstanden werden, die anschließend noch durch den Benutzer korrigiert und ergänzt werden können (diese Form wird manchmal auch als "halbautomatisches" Indexieren, oder, bei MRESSE 1984, als "Mischform" bezeichnet).
Je nach dem bei der Indexierung verwendeten Vokabular kann man weiter differenzieren zwischen der Indexierung mit kontrollierten und mit unkontrollierten Deskriptoren und danach, ob Einwort- oder Mehrwortbegriffe benutzt werden.
Bei kontrollierten Deskriptoren stammen die Indexierungsbegriffe aus einem eingegrenzten Vokabular (d.h. aus einem Thesaurus), während unkontrollierte Deskriptoren im Prinzip alle Begriffe der natürlichen Sprache einschließen.
Bei der Verwendung von Einzelbegriffen, die den Inhalt eines Dokumentes wiedergeben, werden diese später bei einer Suchanfrage evtl. kombiniert (durch logische Verknüpfungen), um die relevanten Dokumente zu finden. Im Unterschied dazu findet bei den Mehrwortbegriffen diese Kombination schon bei der Indexierung statt.
Nach SALTON und MCGILL (1987) werden für die manuelle Indexierung hauptsächlich kontrollierte, zusammengesetzte Deskriptoren verwendet, wohingegen bei automatischen Indexierungsverfahren meist mit Einzelbegriffen gearbeitet wird.

Wie bei der automatischen Indexierung mit unkontrolliertem bzw. kontrolliertem Vokabular vorgegangen werden kann, soll jeweils an einem Beispiel deutlich gemacht werden.

Bei dem ersten Beispiel handelt es sich um ein automatisches Indexierungsverfahren für englischsprachige Texte, beschrieben von SALTON und MCGILL (1987). Im ersten Schritt werden die Einzelbegriffe eines Dokumentes aufgelistet, wobei man sich üblicherweise auf die Begriffe im Titel und Abstract beschränkt. Für diese Einschränkung sprechen einmal die Kosten, die bei der elektronischen Speicherung des gesamten Textes entstehen würden, und die nur unwesentliche Verbesserung der Indexierungsleitung (SALTON und MCGILL 1987: "Werden zusätzlich ganze Dokumenttexte gespeichert, lassen sich im Vergleich zur Indexierung mit Titeln und Abstracts keine großen Verbesserungen mehr erzielen [...]").
Aus den aufgelisteten Begriffen werden dann die sog. "Hochfrequenzbegriffe" entfernt, da sie als Deskriptoren ungeeignet sind. Hochfrequenzbegriffe werden dadurch ermittelt, daß zunächst für jeden Begriff die Häufigkeit in der gesamten Dokumentation (durch Addieren der Häufigkeit dieses Begriffes über alle Dokumente) berechnet wird. Danach werden die Begriffe nach abnehmender Häufigkeit sortiert; alle Begriffe über einem festzulegenden Schwellwert bilden die Hochfrequenzbegriffe. Diese Begriffe (von denen es im Englischen ca. 250 gibt) umfassen bereits 40-50% der Textwörter und können auch in einer Stoppwortliste zusammengefaßt werden.
Zur Ermittlung der Deskriptoren werden dann die verbliebenen Begriffe auf ihren Wortstamm reduziert. Dazu existieren bereits verschiedenene Algorithmen, die immer das längste Suffix (aus einer Suffixliste) entfernen. Diese Algorithmen kann man, unter Berücksichtigung einiger Ausnahmefälle, zur Wortstammgenerierung verwenden.
Danach können die Wortstämme, die als Deskriptoren verwendet werden sollen, mit verschiedenen statistischen Verfahren, auf die hier nicht näher eingegangen werden soll, bestimmt werden. Die ausgewählten Begriffe werden dann (mit oder ohne Gewichtung) den entsprechenden Dokumenten zugeordnet. Bei der gewichteten Indexierung (bei der das Gewicht die Bedeutung des Deskriptors für das jeweilige Dokument wiedergibt) läßt sich diese Zuordnung über einen Dokumentvektor (für jedes Dokument) darstellen. In diesem Dokumentvektor ist für jeden Deskriptor, der insgesamt vergeben wurde, ein Gewicht angegeben (> 0, falls er dem entsprechenden Dokument zugeteilt wurde bzw. = 0, falls er nicht zugeteilt wurde).

Als eines der wenigen Beispiele in der Literatur, die bei der Indexierung auf vorgegebenes Vokabular zurückgreifen, soll das System AIR vorgestellt werden.

Das Verfahren AIR/PHYS wird seit 1987 am Fachinformationszentrum (FIZ) Karlsruhe für die Produktion der Datenbank PHYS eingesetzt. In dieser Datenbank werden wissenschaftliche Veröffentlichungen auf den Gebieten Physik, Astronomie und Astrophysik nachgewiesen (pro Jahr ca. 125.000). Nach KUHLEN (1992) gibt es "[...] weltweit [...] kein zweites automatisches Indexierungssystem, das für eine laufende Datenbankproduktion und für so viele Dokumente angewandt wurde und wird".

Entwickelt wurde dieses System an der TH Darmstadt; die Entwicklungsgeschichte läßt sich (nach KNORZ 1992) kurz zusammenfassen:

- Projekt WAI (1978-1981): Wörterbuchentwicklung für die automatische Indexierung
- Projekt AIR (1981-1983): Weiterentwicklung der automatischen Indexierung und des Information Retrieval
- Projekt AIR/PHYS (1983-1985): Pilotanwendung am Fachinformationszentrum (FIZ) Karlsruhe mit einem neu entwickelten Wörterbuch PHYS/PILOT

Folgende Voraussetzungen kennzeichnen (nach LÜCK et al. 1992) u.a. das AIR/System:

- Die Deskriptoren stammen aus einem verbindlichen Indexierungsvokabular. Dieses Vokabular ist in einem hierarchisch strukturierten Thesaurus mit mehr als 24.000 Begriffen (davon ca. 95 % Mehrwortgruppen) festgelegt.
- Indexiert werden die englischsprachigen Titel und Abstracts der Dokumente.
- Die Indexierung soll automatisch (und nicht nur computerunterstützt) erfolgen.

Die automatische Indexierung mit AIR/PHYS läuft im Prinzip so ab, daß zunächst das Dokument (d.h. Titel und Abstract) in einzelne Wörter zerlegt wird. Aus diesen werden dann grammatische Funktionswörter mit Hilfe einer Stoppwortliste entfernt, alle anderen Wörter auf ihre Grundform reduziert. Physikalische und chemische Formeln werden in standardisierte Formelbezeichnungen umgeformt und Mehrwortgruppen identifiziert. Für jeden Term werden Informationen festgehalten und dann wird geprüft, ob er im Thesaurus mit einem Deskriptor verknüpft ist. Danach werden alle Hinweise auf einen bestimmten Deskriptor in einem Relevanzbeschreibungsvektor festgehalten, auf den dann die Indexierungsfunktion angewandt wird. Als Ergebnis liefert diese Funktion ein Indexierungsgewicht des Deskriptors für das Dokument. Außerdem wird eine weitere Indexierungsfunktion angewandt, die eine gewichtete Indexierung zum Ergebnis hat. Für PHYS werden dann alle Deskriptoren zugeteilt, deren gewichtete Indexierung einen bestimmten Schwellenwert überschreitet.
Die vorgeschlagenen Deskriptoren werden vom Indexierer auf der Grundlage des Volltextes manuell korrigiert.

2.4 Hypertext und Information Retrieval

In den vorhergehenden Abschnitten wurde zum einen Hypertext und zum anderen das Indexieren, eine der Hauptaufgaben des Information Retrieval, beschrieben. Hypertext und Information Retrieval sind eigentlich zwei Gebiete, die voneinander abgegrenzt werden, weil ihnen gegensätzliche Strategien zugrundeliegen: auf der einen Seite das assoziative Browsing, auf der anderen Seite das gezielte Suchen mit einer konkreten Suchanfrage.
Allerdings zeigen sich auch Berührungspunkte zwischen diesen beiden Konzepten, auf die hier kurz eingegangen werden soll. Zu diesen Gemeinsamkeiten bemerkt KUHLEN (1992):

> "So wie sich das klassische Information Retrieval seit Mitte der achtziger Jahre, zumindest in der experimentellen Forschung, zum intelligenten Retrieval entwickelt hat, so deutet sich jetzt zu Beginn der neunziger Jahre eine Symbiose von Hypertext und Information Retrieval an. Der Erfolg dieser Kooperation wird durch den Zuwachs an informationellem Mehrwert bestimmt."

Einer der Berührungspunkte zwischen Hypertext und Information Retrieval wurde bereits in Abschnitt 2.2.3 angesprochen, und zwar die Möglichkeit, Methoden des Information Retrieval (speziell das Indexieren) bei der Generierung von Hyperlinks einzusetzen. KUHLEN (1992) nennt als weitere Berührungspunkte u.a. "die Konversion von Text in Hypertext durch Inhaltserschließungstechniken; die Verwaltung von On-line-Thesauri durch Hypertext [...] ; Hypertext und Freitextretrieval". Nach KRAUSE (1992) werden "Hypertextsysteme als Display- und Browsingwerkzeuge [...] in der Literatur als IIR-Komponenten eingeordnet" (IIR steht für "Intelligentes Information Retrieval", ein relativ neuer Interessenschwerpunkt des Information Retrieval).
Hypertextsysteme können also vom Einsatz von Information-Retrieval-Methoden profitieren, umgekehrt können Information-Retrieval-Systeme durch Hypertext-Techniken angereichert werden.

Im nächsten Abschnitt soll das World Wide Web-Projekt vorgestellt werden; es kann als praktisches Beispiel für die Verbindung der Konzepte Hypertext und Information Retrieval angesehen werden (BERNERS-LEE und CAILLIAU, 1992: "The W3 project merges networked information retrieval and hypertext to make an easy but powerful global information system.").

2.5 Das World Wide Web-Projekt

2.5.1 World Wide Web

World Wide Web (WWW, W3), das am CERN (Centre Européen de Recherches Nucléaires) in Genf entwickelt wurde (s. GRAU 1994), ermöglicht den Zugriff auf bisher bestehende Informationsquellen und Dienste (FTP, WAIS, Gopher, Telnet, EMail und News) im Internet unter einer Benutzeroberfläche, wobei es Hypertext- und Multimedia-Techniken einsetzt.
Es ist schwierig, eine genaue Definition dafür zu finden, was das WWW ist. Je nach Sichtweise wird es bezeichnet als Hypertext-System, Hypertext-Projekt (GILSTER 1994), verteiltes Hypermediasystem (DECEMBER und RANDALL 1994), Informationssystem (MAIER und WILDBERGER 1993), globales Informationssystem (BERNERS-LEE und CAILLIAU) oder als "hypermedia information retrieval initiative" (offizielle Beschreibung). Der erste Projektvorschlag zum WWW vom März 1989 trägt den Titel: "WorldWide Web: Proposal for a HyperText Project" (DECEMBER und RANDALL 1994); das ursprüngliche

Ziel war ein effizienter und einfacher Informationsaustausch zwischen verschiedenen Forschungsgruppen der Hochenergiephysik. Eine spätere Version dieses Projektvorschlages findet sich unter der WWW-Adresse *http://info.cern.ch/hypertext/WWW/Proposal.html*.
An dieser Stelle soll nun auf die Definitin von DECEMBER und RANDALL (1994) zurückgegriffen werden:

> The World Wide Web is a convergence of computational concepts for presenting and linking information dispersed across the internet in an easy accessible way.

WWW arbeitet nach dem Client/Server-Konzept. Die Verbindung zu den Informationsservern wird vom Client über Gateways aufgebaut. Falls es sich (was die Regel ist) um einen WWW-Server handelt, erfolgt die Kommunikation über HTTP (HyperText Transfer Protocol), andernfalls wird die Kommunikation direkt über das Protokoll des jeweiligen Servers abgewickelt. Durch das Präfix der sog. URL (Universal Resource Locator)-Adresse wird festgelegt, über welches Gateway die Kommunikation erfolgt. Eine solche URL-Adresse setzt sich zusammen aus *Gateway://Server-Adresse/Dokument-Pfad/Dateiname*, wobei Gateway z.B. "http" (für eine Datei auf einem WWW-Server) oder "gopher" (für eine Datei auf einem Gopher-Server) sein kann.

Auf die Informationsquellen kann mit einem WWW-Browser (Client) zugegriffen werden, der die Inhalte von Dateien der unterschiedlichsten Formate darstellen kann. Eine solche Software-Schnittstelle zum WWW stellt z.B. das am National Center for Supercomputing Applications an der Universität Illinois entstandene NCSA Mosaic (s.a. 3.3) dar. Dokumente, die angezeigt werden sollen, können zum einen direkt über die URL-Adresse, zum anderen durch Anklicken der Hyperlinks spezifiziert werden. Hyperlinks sind in diesem Fall hervorgehobene Textpassagen oder auch Bilder, hinter denen sich URL-Adressen verbergen, die auf beliebige Quellen im Internet verweisen. Bei Dokumentarten, die nicht direkt verstanden werden, ruft Mosaic externe Programme auf.

Außerdem gibt es die Möglichkeit, über ein "Common Gateway Interface" (CGI), einer standardisierten Schnittstelle, Daten zwischen dem Client und einem beliebigen Server-Programm auszutauschen. Eingegeben werden die Daten über WWW-Formulare (s. 2.5.2) und von dort dem Server-Programm über Environment-Variablen übermittelt.

Sowohl Server- als auch Client-Software gibt es inzwischen für alle gängigen Betriebssysteme, dabei handelt es sich meist um frei verfügbare (bei nicht-kommerziellem Einsatz kostenlose) Software.

Die folgende Tabelle zeigt die Zunahme der WWW-Server. Die Zahlen sind die Ergebnisse eines Programmes, dem "World Wide Web Wanderer"; auf diesem Programm beruht auch die

graphische Darstellung in Abb. 4. Ein anderes Programm fand im August 1994 bereits ca. 7.000 Server. Wenn die Zahlen auch vielleicht ungenau sind, zeigen sie doch den hohen Zuwachs an WWW-Servern in den letzten Jahren.

Datum	*Anzahl der gefundenen Server*
Juni 1993	130
September 1993	204
Oktober 1993	228
November 1993	272
Dezember 1993	623
März 1994	1265
Juni 1994	3184

Abb. 3 Zunahme der WWW-Server (Quelle: DECEMBER und RANDALL 1994)

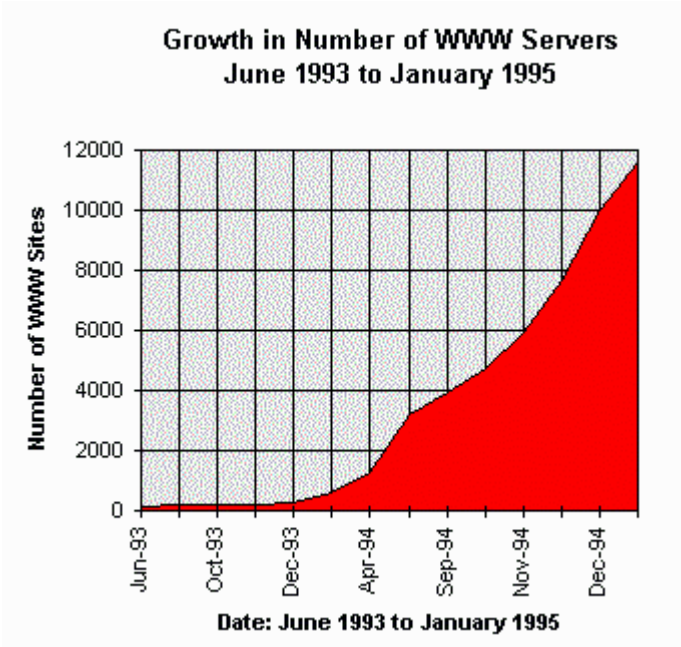


Abb. 4 Zunahme der WWW-Server (Quelle: *http://www.netgen.com/info/growth.html*)

2.5.2 Die Hypertext-Sprache HTML

Speziell für WWW entworfen wurde eine Hypertext-Sprache, die "HyperText Markup Language" (HTML). Abgeleitet ist HTML von SGML (Standard Generalized Markup Language), einer international standardisierten (ISO 8879) Meta-Sprache zum Beschreiben von Markup-Sprachen. Eine bestimmte Markup-Sprache, die unter Verwendung von SGML definiert wird, bezeichnet man als DTD (Document Type Definition); HTML ist also ein "SGML DTD". Verwendet wird HTML seit 1990 im WWW.
HTML-Dateien sind reine ASCII-Dateien; in diese Textdateien werden sog. "Tags" eingefügt, um die Dokumentstruktur festzulegen. So wird z.B. festgelegt, welcher Text eine Überschrift sein soll, wie diese Überschrift dann angezeigt wird, hängt von den Möglichkeiten des jeweiligen Browsers ab.

Im folgenden sollen kurz die wichtigsten HTML-Elemente dargestellt werden und außerdem die sog. "Forms", eine Erweiterung von HTML.

Die meisten der Tags werden paarweise verwendet, z.B. für den Dokumenttitel <TITLE> als Starttag und </TITLE> als Endetag. Die Syntax für das Starttag ist also <*Tagname*>, für das Endetag </*Tagname*>.

Header

Innerhalb des Headers werden allgemeine Informationen über das Dokument abgelegt, die nicht vom Browser angezeigt werden. Der Anfang des Headers wird mit <HEAD> gekennzeichnet, das Ende mit </HEAD>. Eine Information, die z.B. innerhalb des Headers definiert wird, ist der Dokumenttitel.

Titel

Der Titel eines Dokumentes wird festgelegt durch

<TITLE>Dokumenttitel</TITLE>

Die folgenden HTML-Elemente werden innerhalb des eigentlichen Dokumenttextes benutzt, der durch <BODY> und </BODY> geklammert ist. Sie müssen in derjenigen Reihenfolge verwendet werden, in der sie später im Dokument erscheinen sollen.

Überschriften

In HTML sind sechs Überschriftengrößen verfügbar, numeriert von 1 (größte Schrift) bis 6 (kleinste Schrift). Die Syntax für das "heading-tag" lautet wie folgt:

<Hx>Text der Überschrift</Hx> $x \in \{1,2,3,4,5,6\}$

Bei Überschriften wird außerdem automatisch ein Absatz erzeugt.

Absätze

Leerzeilen bzw. Zeilenvorschübe im Sourcetext werden ignoriert, deshalb ist es notwendig, bei gewünschten Absätzen explizit ein <P> anzugeben.

Bilder

Innerhalb von HTML-Dokumenten können Bilder mit dem IMG-Element angezeigt werden. Dabei gibt das SRC-Attribut an, welches Bild eingebunden wird, das ALIGN-Attribut legt fest, ob der zugehörige Text an den oberen (ALIGN=TOP) oder unteren Bildrand (ALIGN=BOTTOM, Default-Wert) bzw. an die Bildmitte (ALIGN=MIDDLE) gesetzt wird.

Bsp.: <IMG ALIGN=MIDDLE SRC="RESEDA/gif/pamtwama.gif">

Verweise

Ein Anker (anchor) wird dazu genutzt, um aus dem aktuellen Dokument auf ein anderes Hypertext-Dokument oder z.B. auch auf ein Bild zu verweisen. Zwischen <A> und </A> steht der Text, der im Dokument als Hypertextverweis erscheinen soll. Hypertextverweise sind in NCSA Mosaic farbig und unterstrichen dargestellt. Anstelle des Textes (bzw. zusätzlich) ist auch das IMG-Element erlaubt, das als Hypertextverweis dann farbig umrahmt erscheint.
Der Name des Dokumentes, zu dem verwiesen werden soll, wird mit dem Attribut HREF festgelegt. Zusätzlich kann man mit "*#name*" zu einer bestimmten Stelle in diesem Dokument springen, die durch das Attribut NAME=*name* gekennzeichnet sein muß (diese Stelle kann ebenfalls wieder ein Verweis sein). "*#name*" kann auch auf eine andere Stelle im gleichen Dokument verweisen, was besonders bei längeren HTML-Seiten sinnvoll sein kann.
Ist das Dokument, auf das mit HREF verwiesen wird, keine HTML-Seite, wird es (in Abhängigkeit von der Dateierweiterung) mit einem externen Programm angezeigt, bei GIF-Bildern z.B. mit "xv".

Bsp.: in Datei X: <A HREF="/Umweltdaten/thes/Y.html#siedlung">Siedlung</A>
in Datei Y: <A NAME="siedlung">Siedlung</A>

In diesem Beispiel gelangt man durch Anklicken des Wortes "Siedlung" in der Datei X zur Datei Y und dort an die Stelle, die durch NAME="siedlung" gekennzeichnet ist.

<A HREF="/Umweltdaten/abb1.gif">Abb. 1</A>

Hier wird bei Anklicken von "Abb. 1" das Bild "abb.1.gif" mit dem "xv" angezeigt.

Nichtnumerierte Listen

Nichtnumerierte Listen beginnen mit dem Starttag <UL>, das Endetag ist </UL>. Dazwischen stehen die einzelnen Einträge, alle mit <LI> beginnend. Diese Einträge erscheinen mit vorangestelltem "•".

Bsp.: Dokumente:<P>
<UL>
<LI>Ökologisches Wirkungskataster
<LI>Ökologisches Datenbanksystem
<LI>Biotopkartierung
<LI>Gebiets- und Biotopschutz
<LI>Waldschutzsituation
<LI>Orkanschäden 1990
<LI>Imissionsökologische Waldzustandserhebung
<LI>Waldbiotopkartierung
</UL>

wird dargestellt als

Dokumente:

- Ökologisches Wirkungskataster
- Ökologisches Datenbanksystem
- Biotopkartierung
- Gebiets- und Biotopschutz
- Waldschutzsituation
- Orkanschäden 1990
- Imissionsökologische Waldzustandserhebung
- Waldbiotopkartierung

Beschreibungslisten

Zwischen Start- und Endetag einer Beschreibungsliste (<DL> bzw. </DL>) steht eine Folge von Beschreibungstiteln (gekennzeichnet durch <DT>) mit den dazugehörigen Beschreibungen (gekennzeichnet durch <DD>). <DT> und <DD> treten also immer paarweise auf.

Bsp.: <DL>
<DT>Agrarflurbereinigung:
<DD>Ziel jeder Flurbereinigung ist es, neben der Verbesserung der Agrarstruktur zur Erhaltung und Gestaltung der Kulturlandschaft beizutragen. Sowohl die landschaftsökologischen Vorstellungen als auch die landwirtschaftlichen Anforderungen werden in einem flächendeckenden, weiträumigen Gesamtkonzept berücksichtigt.
<DT>Biotopvernetzung:
<DD>Biotopvernetzung ist ein Konzept zur ökologischen Gestaltung der Kulturlandschaft. In Gebieten mit intensiver landwirtschaftlicher Nutzung sollen Biotope und Ausgleichsräume zur Verbesserung der Lebensbedingungen der Tier- und Planzenwelt erhalten, ergänzt, geschaffen und miteinander verbunden werden.
<DT>Landschaftsplanung:
<DD>Die Landschaftsplanung liefert als Planungsinstrument den ökologischen Orientierungsrahmen für die weitere Entwicklung von Natur und Landschaft.
</DL>

wird dargestellt als

Agrarflurbereinigung:
: Ziel jeder Flurbereinigung ist es, neben der Verbesserung der Agrarstruktur zur Erhaltung und Gestaltung der Kulturlandschaft beizutragen. Sowohl die landschaftsökologischen Vorstellungen als auch die landwirtschaftlichen Anforderungen werden in einem flächendeckenden, weiträumigen Gesamtkonzept berücksichtigt.

Biotopvernetzung:
: Biotopvernetzung ist ein Konzept zur ökologischen Gestaltung der Kulturlandschaft. In Gebieten mit intensiver landwirtschaftlicher Nutzung sollen Biotope und Ausgleichsräume zur Verbesserung der Lebensbedingungen der Tier- und Planzenwelt erhalten, ergänzt, geschaffen und miteinander verbunden werden.

Landschaftsplanung:

Die Landschaftsplanung liefert als Planungsinstrument den ökologischen Orientierungsrahmen für die weitere Entwicklung von Natur und Landschaft.

Formatierungen

In HTML stehen verschiedene logische und physische Schriftarten zum Hervorheben von Textstellen zur Verfügung.
Physische Schriftarten können z.B. mit <I> und </I> (für kursiv) oder <B> und </B> (für fett) gekennzeichnet werden, logische Schriftarten legen dagegen nicht das Aussehen, sondern die Bedeutung fest. Sie entsprechen damit eher der "Philosophie" von HTML bzw. SGML und sollten wenn möglich verwendet werden.
Folgende logische Schriftarten können u.a. verwendet werden:

<CITE>Text</CITE>	für Zitate (in NCSA Mosaic: *Text*)
<STRONG>Text</STRONG>	zum Hervorheben (in NCSA Mosaic: **Text**)
<CODE>Text</CODE>	für Computercode (in NCSA Mosaic: T e x t)

Eine Erweiterung der HTML-Syntax stellen die HTML-Formulare ("Forms") dar. Diese dienen dazu, Daten vom Client an den Server zu übermitteln. Die in ein Formular eingegebenen Werte können dann als Eingabewerte für ein beliebiges Programm (CGI-Skript) genutzt werden. Die Ausgabewerte dieses Programmes (z.B. eine HTML-Seite) werden wieder an den Client zurückgegeben.

Die einzelnen Elemente des Formulars werden innerhalb von <FORM> und </FORM> festgelegt, zusätzlich zu den oben beschriebenen HTML-Elementen können hier weitere HTML-Elemente (z.B. Texteingabefelder) verwendet werden. Attribute des Form-Tags sind "ACTION" und "METHOD", wobei durch "ACTION" über eine URL-Adresse das Skript (Programm) festgelegt wird, zu dem die Daten übermittelt werden, und durch "METHOD" die Art der Datenübermittlung.

Bsp.: <FORM METHOD="POST"
ACTION="http://faw.uni-ulm.de:9876/cgi-bin/htgrep.csh">

Durch METHOD="POST" wird festgelegt, daß die Eingabewerte über die Standardausgabe an das Shell-Skript "htgrep.csh" übergeben werden.

Das wesentliche Element innerhalb eines Formulars, das für die meisten Anwendungen genügt, ist das INPUT-Tag, mit dem Eingabefelder definiert werden können. Wichtige

Attribute sind hier "TYPE" (legt den Typ des Eingabeelementes fest), "NAME" (legt den Variablennamen in der Datenstruktur fest, die zum Server-Programm gesendet wird) und "VALUE" (zur Festlegung von Default-Werten). Gültige Werte für "TYPE" sind:

- TEXT für die Eingabe eines alphanumerischen Strings
- RADIO zur Auswahl genau eines Buttons von mehreren
- SUBMIT zur Übermittlung der Eingabewerte an das Server-Programm
- RESET zum Zurücksetzen aller Eingabefelder.

Bsp.: <INPUT TYPE="SUBMIT" VALUE="SEARCH">

definiert einen "Submit"-Button mit der Aufschrift "SEARCH".

Neben dem INPUT-Tag gibt es noch das SELECT-Tag (für Auswahllisten) und "TEXTAREA" (für die Eingabe von längerem Text).

3. Umweltberichte als Hypertext

3.1 Aufgabenspezifikation

Der Bericht "Umweltdaten 91/92" (s. 3.4.1) sowie der Umwelt-Thesaurus des Umweltbundesamtes (s. 3.4.2) sollen für den Endbenutzer als Hypertext im WWW zur Verfügung gestellt werden. Dabei soll der Zugang zum Umweltbericht über Inhaltsverzeichnis, alphabetischen Index und Thesaurus möglich sein. Zu diesem Zweck müssen Programme erstellt werden, die sowohl den Umweltbericht als auch den Thesaurus als Hypertextseiten für die Verwendung im WWW aufbereiten.

An diese Hypertextseiten bzw. die zu entwickelnden Programme werden folgende EDV-technischen Anforderungen gestellt:

1. Die Hypertextseiten sollen auf einer Sun-Workstation unter Unix getestet und für die Installation auf einem WWW-Server unter dem Betriebssystem Unix vorbereitet werden.

2. Alle Hypertextseiten sollen sich im Filesystem eines einzigen Server-Rechners befinden und dort einen Teilbaum des Dateiverzeichnisses belegen.

3. Hyperlinks sollen relativ zur Wurzel dieses Teilbaums adressiert sein, so daß sich dieser Teilbaum aller Hypertextseiten und Unterverzeichnisse in seiner Gesamtheit auf einen anderen Server-Rechner kopieren läßt, ohne daß Hyperlink-Referenzen nacheditiert werden müssen.

4. Die Hypertextseiten dürfen keine Hyperlinks auf externe Rechner enthalten.

5. Der Abruf der Hypertextseiten soll mit Hilfe der NCSA-Mosaic-Software auf einer Vielzahl von Clientrechnern, insbesondere auch VAX/VMS, möglich sein, sofern eine HTTP-Verbindung zwischen Client- und Server-Rechner besteht.

6. Die Programme zur Erstellung der Dokumente sowie ein CGI-Skript für die Volltextrecherche zur Laufzeit sollen auf einer Sun-Workstation compiliert und getestet werden.

3.2 Das System aus Benutzer- und aus Autorensicht

In diesem Kapitel soll zum einen gezeigt werden, wie sich der erstellte Hypertext für den Benutzer präsentiert, und zum anderen, wie sich das System aus der Sicht des Autors, der den Hypertext generiert, darstellt.

Für den Benutzer gibt es verschiedene Möglichkeiten, auf die Informationen im Umweltbericht zuzugreifen.
Eine erste Möglichkeit, die den Thesaurus noch nicht einbezieht, ist die Navigation über die hierarchische Kapitelstruktur des Berichtes, d.h. ausgehend vom Gesamtinhaltsverzeichnis kann zu den Inhaltsverzeichnissen der einzelnen Kapitel und von dort zu den Berichtsabschnitten gesprungen werden.

Ein weiterer Einstiegspunkt ist über den alphabetischen Index, der Deskriptoren und Synonymbegriffe enthält, gegeben. Von den Begriffen des alphabetischen Index kann direkt auf das Deskriptordokument des entsprechenden Deskriptors zugegriffen werden. Von diesem Deskriptor wird auf Ober- und Unterbegriffe des Deskriptors verwiesen; auf diese Weise sind die Deskriptordokumente ebenfalls hierarchisch untereinander verknüpft. Außerdem ist hier die Verbindung zwischen der Berichtsstruktur und dem Thesaurus gegeben dadurch, daß eine Deskriptorseite mit allen Berichtsabschnitten verknüpft ist, denen dieser Deskriptor zugeordnet wurde.

Die Navigationsmöglichkeiten sollen anhand einiger Beispielseiten verdeutlicht werden. Die erste Abbildung zeigt die Einstiegsseite, die man unter der URL-Adresse *http://faw.uni-ulm.de:9876/Umweltdaten/start.html* erreicht.

Abb. 5 Startseite

Von hier kann man direkt zu den Berichtsabschnitten (durch Anklicken von "Umweltdaten 91/92"), zum alphabetischen Thesaurus oder zur Volltextrecherche gelangen. Folgt man den "Umweltdaten 91/92", erreicht man eine weitere Übersichtsseite und von dort das Gesamtinhaltsverzeichnis der Umweltdaten (Abb. 6).

Abb. 6 Gesamtinhaltsverzeichnis der Umweltdaten

Jeder Eintrag in diesem Gesamtinhaltsverzeichnis führt zu einem Inhaltsverzeichnis für die einzelnen Kapitel, z.B. dem Inhaltsverzeichnis für das Kapitel "Abfall, Altlasten" (Abb. 7).

Abb. 7 Inhaltsverzeichnis des Kapitels "Abfall, Altlasten"

Von hier aus gelangt man zu den eigentlichen Umweltdaten, z.B. zum Berichtsabschnitt "Altlasten" (Abb. 8).

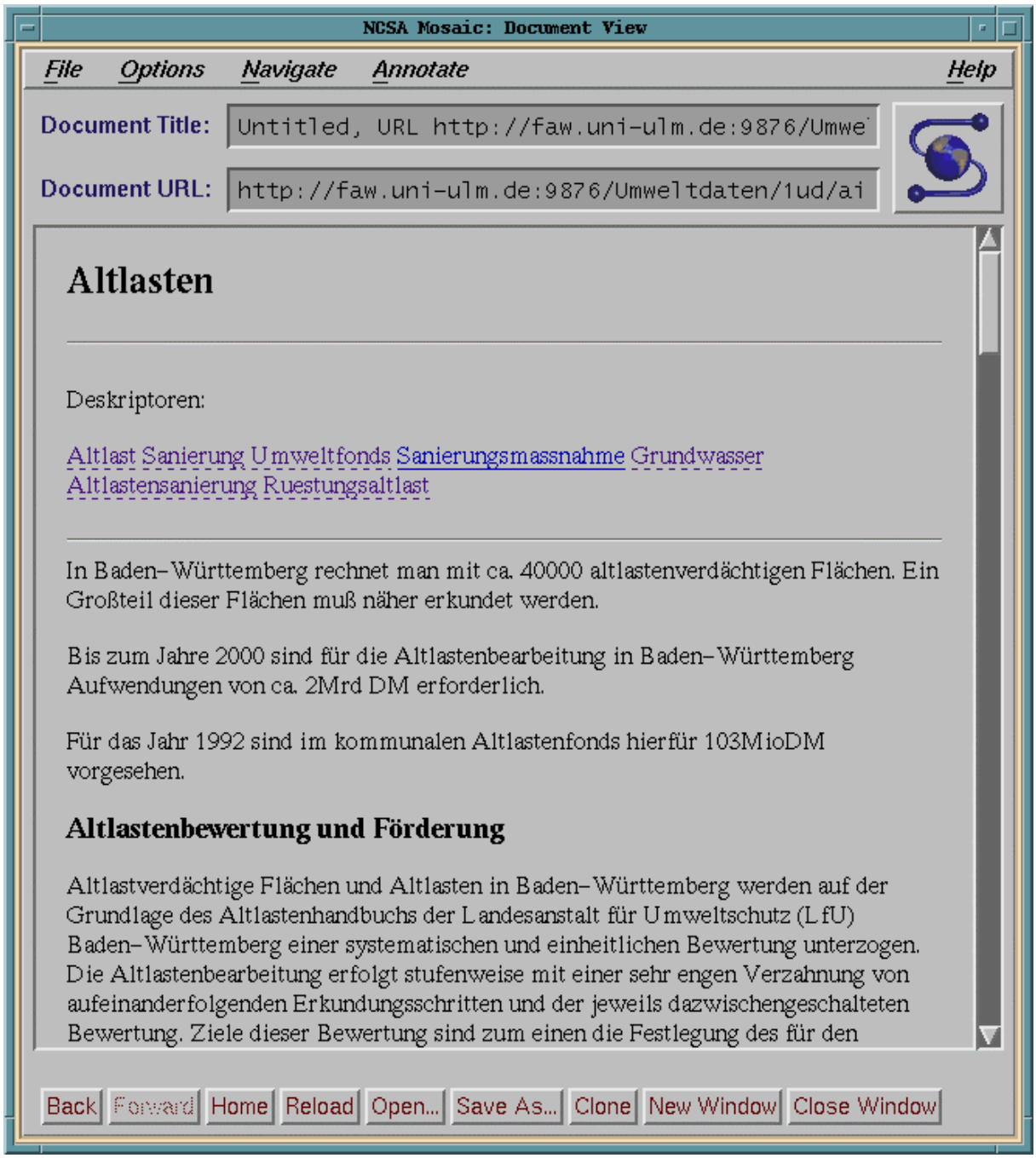


Abb. 8 Berichtsabschnitt "Altlasten" (oberer Teil)

Im oberen Teil des Berichtsabschnittes sind die vergebenen Deskriptoren zu sehen, im unteren Teil die Hyperlinks zu Tabellen und Abbildungen (Abb. 9). Außerdem sind Hyperlinks enthalten zurück zur Eingangsseite und vorwärts zum nächsten Kapitel. Der Hyperlink zum nächsten Kapitel ermöglicht ein schrittweises Durchblättern des gesamten Umweltberichtes, so daß man beim Folgen diese Pfades alle Berichtsabschnitte auf jeden Fall einmal erreicht.

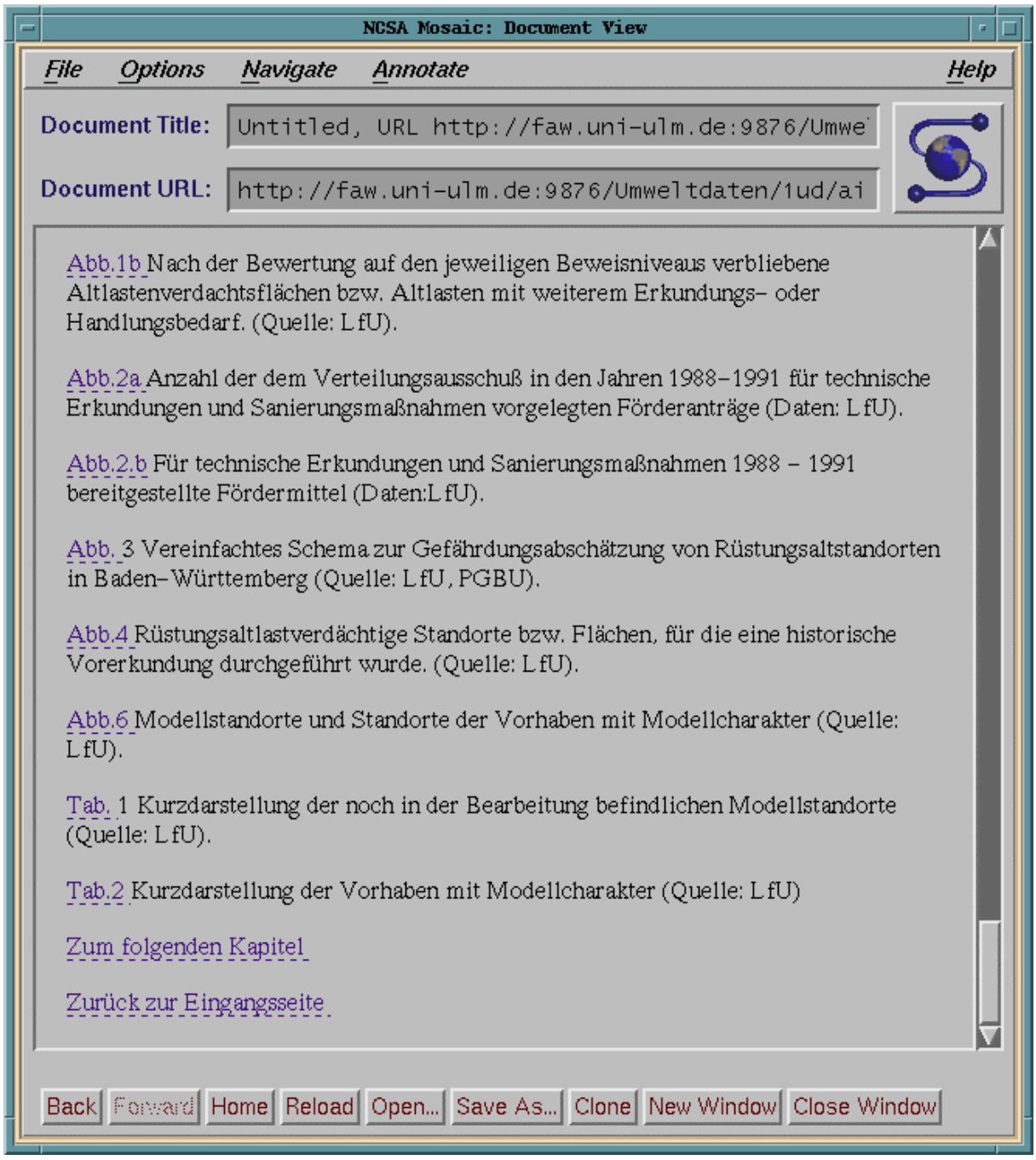


Abb. 9 Berichtsabschnitt "Altlasten" (unterer Teil)

Durch Anklicken von beispielsweise "Abb. 6" erhält man eine Karte (Abb. 10), bei Anklicken von "Tab. 1" eine Tabelle (Abb. 11).

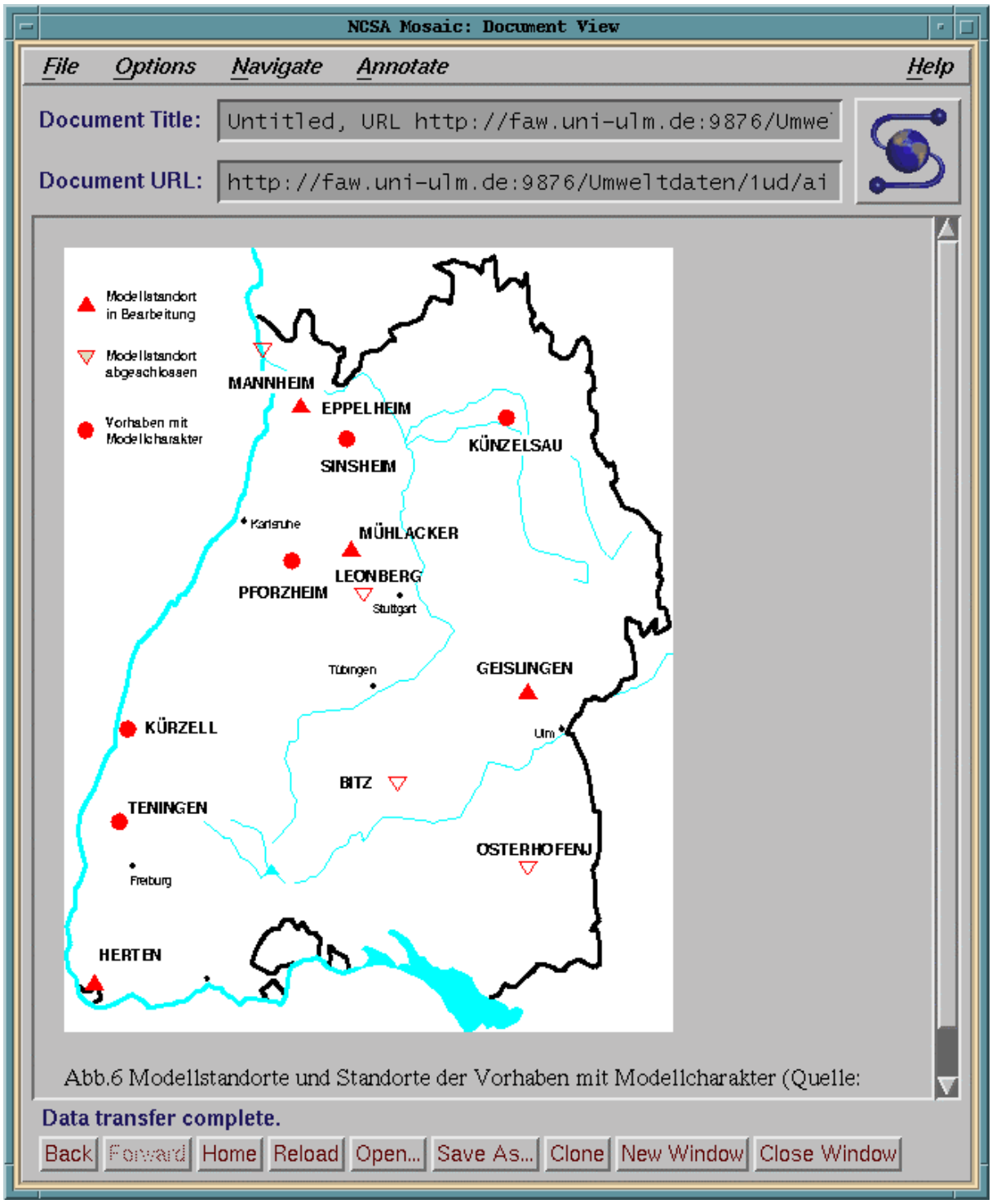


Abb. 10 Karte zu "Modellstandorten"

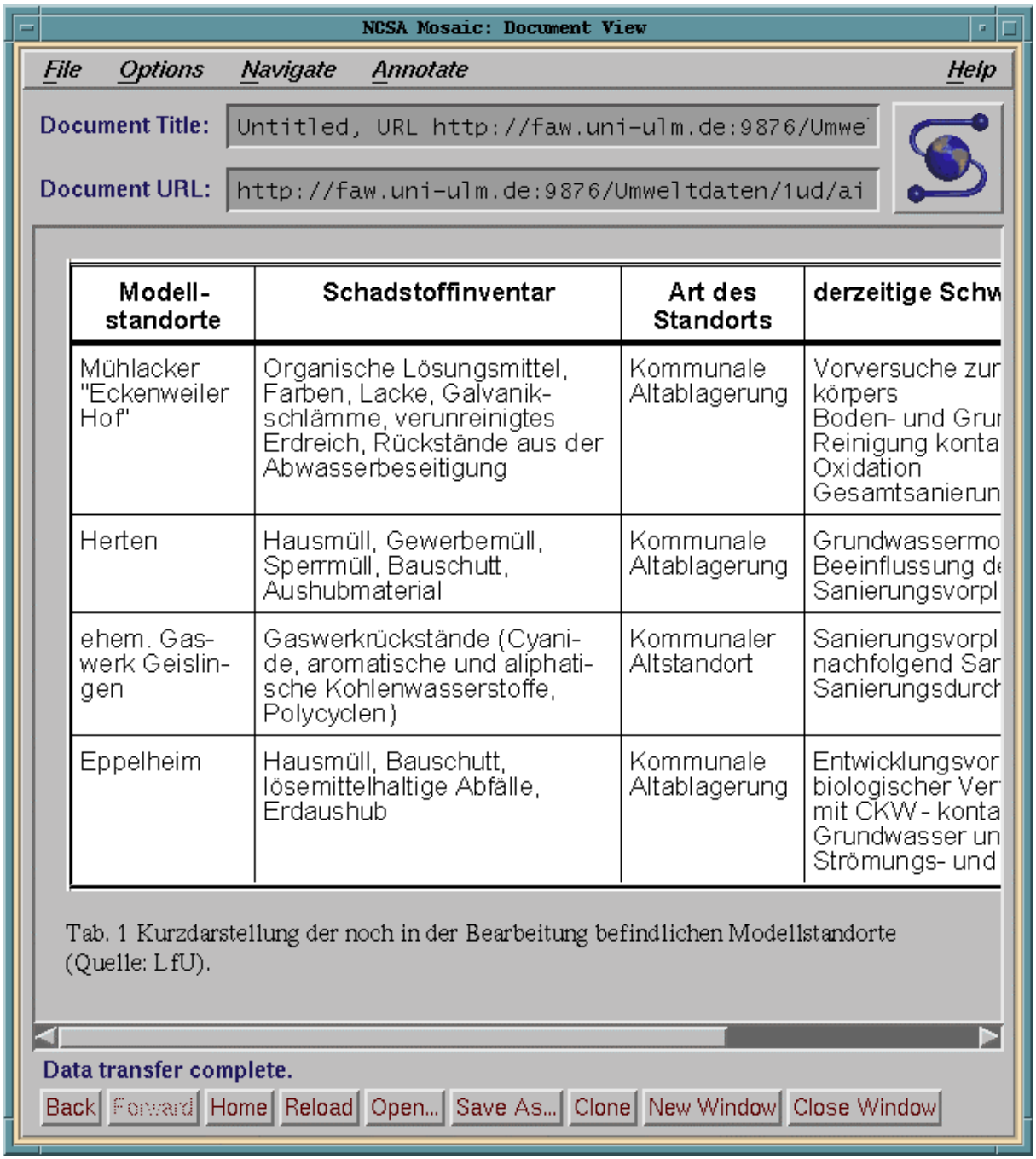


Abb. 11 Tabelle zu "Modellstandorten"

Der Übergang zum Thesaurus ist durch Anklicken der Deskriptoren möglich. Wählt man den Deskriptor "Grundwasser", kommt man zum entsprechenden Deskriptordokument (Abb. 12), das Verweise zu allen weiteren Dokumente mit diesem Deskriptor enthält.

Abb. 12 Deskriptordokument für "Grundwasser"

Der Deskriptor selbst verweist auf den alphabetischen Index (Abb. 13), und zwar direkt zum Eintrag "Grundwasser".

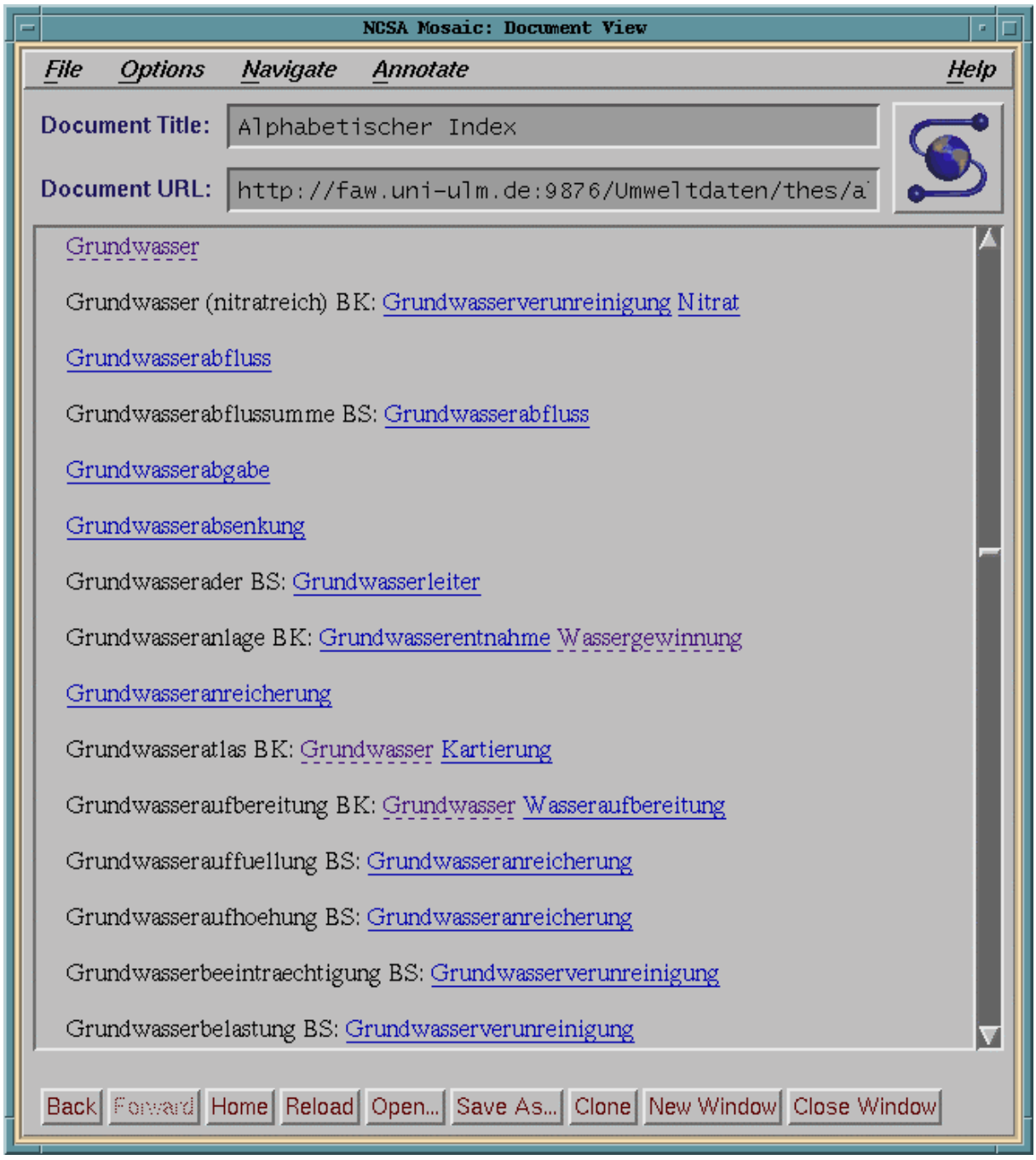


Abb. 13 Alphabetischer Index

Weitere Navigationsmöglichkeiten sind im Deskriptordokument durch die Ober- und Unterbegriffe gegeben. Klickt man auf den Unterbegriff "Grundwasserbeschaffenheit", gelangt man zum Deskriptordokument für diesen Begriff (Abb. 14).

Abb. 14 Deskriptordokument für "Grundwasserbeschaffenheit"

Von den Deskriptordokumenten kann man wieder zu den Berichtsabschnitten gelangen durch Auswählen der Dokumente, die ebenfalls diesen Deskriptor aufweisen. Zum Deskriptor "Grundwasserbeschaffenheit" gibt es nur einen Berichtsabschnitt mit dem Titel "Grundwasserbeschaffenheit" (Abb. 15).

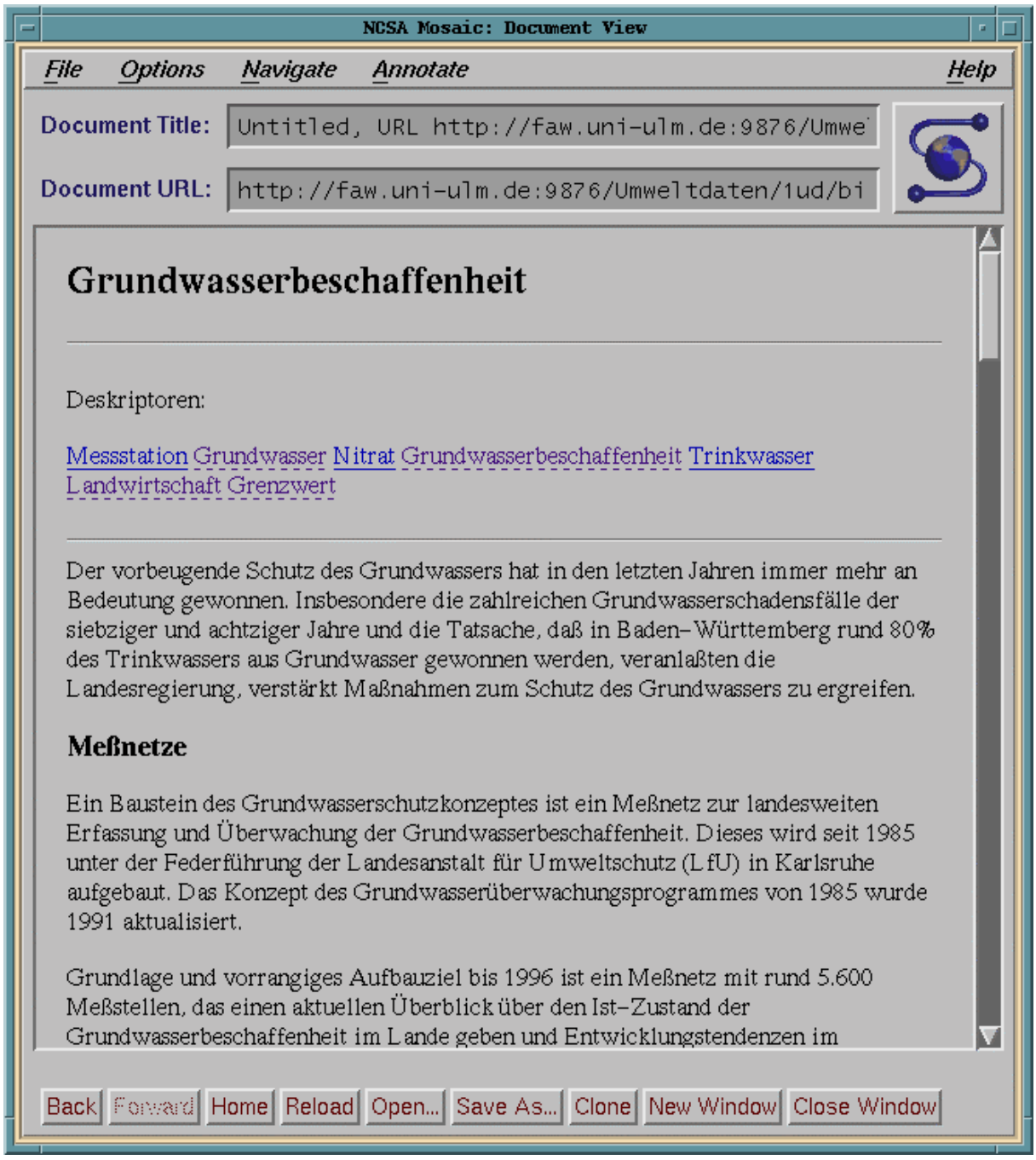


Abb. 15 Berichtsabschnitt "Grundwasserbeschaffenheit"

Zu den Berichtsabschnitten kann man auch über die Volltextsuche gelangen. Dazu klickt man auf der Startseite "Volltextsuche" an, man erhält dann ein Formular zum Eingeben der Suchbegriffe (Abb. 16).

Abb. 16 Formular zum Eingeben der Suchbegriffe

Als Ergebnis erhält man eine Seite, die Hyperlinks zu allen Berichtsabschnitten enthält, für die die Suchbedingung erfüllt ist.

Welche Schritte sind nun nötig, um den oben beschriebenen Hypertext zu erstellen ? Als erstes müssen allen Berichtsabschnitten Deskriptoren aus dem Thesaurus zugeordnet werden. Dazu muß ein Shell-Skript ("indexALLE.csh") aufgerufen werden, das alle Berichtsabschnitte automatisch indexiert. Danach können diese vergebenen Deskriptoren vom Autor manuell korrigiert werden, indem die Berichtsabschnitte einzeln nacheditiert werden. Anschließend wird das Shell-Skript "deskriptDEalle.csh" aufgerufen, welches in die Deskriptordokumente die Hyperlinks zu den Berichtsabschnitten, in denen der jeweilige Deskriptor vorkommt, einfügt.

3.3 Entwicklungsumgebung

Die Programme wurden auf einer Sun-Workstation unter dem Betriebssystem Unix erstellt und ausgeführt.
Verwendet wurden die Programmiersprachen "C" (mit dem Compiler "gcc") und "nawk". Bei "nawk" handelt es sich um eine erweiterte Version von "awk", einer Programmiersprache, mit der Dateien auf bestimmte Textmuster untersucht und entsprechende Aktionen veranlaßt werden können. Außerdem wurden Shell-Skripts erstellt, die u.a. auf folgende Unix-Tools zurückgreifen:

- find: durchsucht anzugebende Verzeichnisse rekursiv nach Dateien mit bestimmten Eigenschaften
- grep: durchsucht eine Datei nach Zeichenketten oder regulären Ausdrücken
- sed : nicht-interaktiver Zeileneditor ("stream editor")
- sort: sortiert Dateien
- tr: ersetzt oder löscht Zeichen.

Der Abruf der erstellten Hypertextseiten ist mit der Software "NCSA Mosaic" möglich. Dieser am National Center for Supercomputing Applications (NCSA) entwickelte Browser ist die wohl derzeit bekannteste Software-Schnittstelle zum WWW. Verfügbar sind Versionen für das X Window System, für Apple Macintosh und für Microsoft Windows. Die folgenden Angaben beziehen sich auf die Mosaic-Version für X Windows (kurz: X Mosaic). Abb. 17 zeigt die Oberfläche von X Mosaic; das Einstiegsdokument (die Homepage) kann individuell festgelegt werden. Von X Mosaic werden, neben dem Anzeigen der HTML-Dokumente, u.a. folgende Funktionen bereitgestellt:

- Neben der Navigationsmöglichkeit über das Anklicken von Textstellen oder der direkten Angabe einer URL-Adresse können Dokumente über die "Hotlist" und die "Window History" ausgewählt werden. Bei der "Hotlist" hat man selbst die Möglichkeit, interessante Dokumente hinzuzufügen, so daß man diese später leicht

wiederfinden kann. In der "Window History" sind alle Titel der Dokumente, die während der laufenden Sitzung angezeigt wurden, aufgelistet.

- Jedes Dokument kann mit Anmerkungen versehen werden. Diese Annotationen erscheinen als Hyperlinks mit Namens- und Datumsangabe am Ende des Dokuments und können nachträglich editiert bzw. gelöscht werden.

- Beim Speichern, Drucken oder Mailen von Dokumenten stehen verschiedene Datei-Formate zur Auswahl: formatierter Text, unformatierter Text, HTML und PostScript.

- Angezeigt werden können nicht nur HTML-Dateien, sondern auch andere Formate, die dann mit externen Programmen angezeigt werden, z.B. GIF-Dateien mit dem "xv" und PostScript-Dateien mit "ghostview".

Abb. 17 Oberfläche von Mosaic für das X Window System

3.4 Eingabedaten

3.4.1 Umweltdaten

Bei den verwendeten Umweltdaten handelt es sich um einen Situationsbericht zur Umwelt in Baden-Württemberg. Dieser Bericht, herausgegeben vom Umweltministerium Baden-Württemberg in Zusammenarbeit mit der Landesanstalt für Umweltschutz (LfU), liegt in gedruckter Form ("Umweltdaten 91/92") sowie als elektronisches Dokument im Format des "Corel Ventura Publisher" (Version 4) vor. In "Quark Express" liegt ein ähnlich aufgebauter Bericht des Umweltbundesamtes (UBA) vor; für das GLOBUS-Projekt wurde dieser jedoch nicht rechtzeitig zur Verfügung gestellt.

Für die Verwendung in dieser Arbeit wurden die Dokumente in HTML-Format konvertiert zur Verfügung gestellt. Für viele Markup-Sprachen (wie z.B. LaTeX) existieren bereits Konverter, die auch automatisch Hyperlinks z.B. von Inhaltsverzeichnissen zu den Kapiteln generieren; Konverter, die Textverarbeitungsdateien umwandeln, setzen jedoch meist nur die Syntax um. Einen Überblick über existierende Konverter findet sich im WWW unter der Adresse *http://info.cern.ch/hypertext/WWW/Tools/Filters.html*.
Für die Umsetzung der Ventura-Publisher- bzw. Quark-Express-Dokumente wurde vom IPF ein eigener Konverter entwickelt, der die Texte in HTML-Format umsetzt und gleichzeitig Hyperlinks von den Inhaltsverzeichnissen zu den einzelnen Kapiteln sowie von den Abbildungsunterschriften zu den Abbildungen generiert. Die Graphiken des Umweltberichtes wurden ebenfalls am IPF für die Verwendung im WWW aufbereitet. Dazu war es nötig, aus den Graphikdateien Ventura-spezifische Steuerzeichen von Hand zu entfernen, und anschließend nicht im GIF vorliegende Dateien (sondern z.B. in EPS, WMF oder Lotus-Format) mit Graphik-Konvertierungsprogrammen in GIF umzuwandeln.

3.4.2 Thesaurus

Für die automatische Indexierung stand der Umwelt-Thesaurus des Umweltbundesamtes zur Verfügung. Da dieser Thesaurus speziell für die Datenbanken ULIDAT und UFORDAT erstellt und fortgeschrieben wird, ist das Thema dieser Datenbanken auch für den Inhalt des Umwelt-Thesaurus bestimmend. Thematischer Schwerpunkt dieser Datenbanken "... ist die Beeinflussung der Umwelt des Menschen durch menschliches Handeln, die Belastungsfaktoren, Auswirkungen, Rückwirkungen auf den Menschen und Abwehrmaßnahmen (Umweltschutz). Eingeschlossen ist auch die Nutzung von Rohstoffen und Energie." (UMWELTBUNDESAMT 1993).

Der Thesaurus enthält eine begrenzte Liste von der natürlichen Sprache entnommenen Bezeichnungen, die es ermöglichen sollen, die Inhalte eines Fachgebietes mit wenigen einheitlichen Bezeichnungen (Stichworte, Schlagwörter, Deskriptoren) zu charakterisieren. Diese Deskriptoren stehen in hierarchischer Beziehung zueinander, d.h. zu den Begriffen existieren Verweise auf Ober- und Unterbegriffe.

Zusätzlich zu den Deskriptoren sind in einem weiteren Teil des Thesaurus abweichende Benennungen (Synonyme) aufgeführt, mit einem Hinweis auf die statt dieser Benennungen zu benutzenden Deskriptoren. Bei einer speziellen Form der Synonymumsetzung sind zwei oder mehr Deskriptoren angegeben (Kombination), d.h. ein Begriff muß durch mehrere Deskriptoren umschrieben werden.

Der Thesaurus steht in unterschiedlichen Dateien (im ASCII-Format) zur Verfügung:

- als alphabetisch sortierte Liste
  Diese Liste enthält alle Deskriptoren in alphabetischer Reihenfolge, jedoch keine Synonym- oder Hierarchieinformationen.

- als hierarchisch strukturierte Liste
  Alphabetisch aufgelistet sind alle obersten Oberbegriffe (d.h. Begriffe, die selbst keinen Oberbegriff mehr haben) mit allen Unterbegriffen und Verzweigungen. Die Hierarchiebeziehungen sind hierbei durch entsprechende Einrückungen der Begriffe kenntlich gemacht.

- als systematische Liste
  Hier sind alle Deskriptoren mit weiteren Informationen (Ober- und Unterbegriffe, verwandte Begriffe,...) sowie Synonyme mit Deskriptoren aufgeführt.

Da für alle Programme der systematische Thesaurus (in verschiedenen abgewandelten Formen) als Grundform benutzt wird, soll der Aufbau dieser Datei mit ihren Abkürzungen kurz dargestellt werden.
Alle Deskriptoren und Synonyme sind alphabetisch aufgelistet (ein Begriff pro Zeile). Zu den Deskriptoren sind weiterhin, durch die aufgeführten Abkürzungen gekennzeichnet, folgende Informationen angegeben:

- Oberster Oberbegriff (OOB)
  Hier findet man Deskriptoren, die keinen weiteren Oberbegriff mehr haben, also an der Spitze des gesamten Hierarchiebaumes stehen.
- Oberbegriff (OB)
  Nach "OB" sind die nächst-erweiternden Deskriptoren angegeben.

- Unterbegriff (UB)
  Nach "UB" sind die nächst-einengenden Deskriptoren aufgelistet.
- Benutzt für (BF)
  Dahinter sind Synonyme, also Nicht-Deskriptoren, des voranstehenden Deskriptors angegeben.
- Verwandter Begriff (VB)
  Unter "VB" sind thematisch verwandte, jedoch hierarchisch nicht verknüpfbare Begriffe aufeinander bezogen.
- Englische Übersetzung (Eng)
  Hier ist eine englische Übersetzung des Deskriptors angegeben.
- Erläuterung (Erl)
  Die Erläuterung enthält Definitionen sowie Hinweise zur Anwendung und zum "Aufsetzpunkt" in parallelen Mikrothesauri (für sehr spezielle Fragestellungen).

Bei den Synonymen kann man folgende Informationen finden:

- Benutze Synonym (BS)
  Dahinter ist der Deskriptor angegeben, der statt des Synonyms benutzt werden soll.
- Benutze Kombination (BK)
  Nach "BK" ist eine Deskriptorkombination aufgelistete, die an Stelle des Synonyms verwendet werden soll.

Anzumerken ist, daß bei den Deskriptoren meist nur ein Teil der oben aufgeführten Informationen zu finden ist; bei den Synonymen ist entweder "BS" oder "BK" angegeben.

Nachfolgend zur Verdeutlichung ein kurzer Auszug aus der systematischen Liste:

[...]

Umweltaufklaerung
BS Umwelterziehung
Umweltauflage
BS Umweltschutzauflage
Umweltaufwendung
BS Umweltschutzkosten
Umweltausgabe
BS Umweltschutzkosten
Umweltauswirkung
UB Umweltbeeintraechtigung
Oekologische Wirksamkeit
Wirkungsforschung
Umweltbeauftragte
BS Umweltschutzbeauftragter
Umweltbedrohung
BS Umweltgefaehrdung
Umweltbeeintraechtigung
OOB Umweltauswirkung

OB Umweltauswirkung
UB Umweltgefaehrdung
Umweltschaden
Umweltveraenderung
Umweltverschmutzung
Umweltzerstoerung
BF Umweltprobleme (global)
Umweltverschmutzung (global)
Eng environmental infrigement
environmental disorder
impairement of the environment
impairement by environmental influence
Umweltbehoerde
OOB Oeffentliche Verwaltung
Interessenverband
Soziale Bewegung
OB Behoerde
Umweltschutzorganisation
UB Europaeische Umweltagentur
Naturschutzbehoerde
BF Umweltamt
Umweltschutzaemter
Umweltschutzamt
Umweltschutzbehoerde
Umweltverwaltung
MURL
Eng environmental authority
environmental protection agency
Umweltbelastbarkeit
OOB Belastbarkeit
OB Belastbarkeit
UB Tragfaehigkeit (oekologisch)
Eng pollution absorption capacity
pollution loading capacity
pollution carrying capacity
[...]

3.5 Systementwurf

Für die Darstellung des Gesamtsystems wird ein Datenflußdiagramm (DFD) verwendet, wie es bei YOURDON (1992) beschrieben ist. Ein solches DFD besteht aus den Komponenten "Prozeß", "Fluß", "Speicher" und "Terminator". Die Bedeutung der einzelnen Komponenten und die für sie verwendeten Symbole können der folgenden Tabelle entnommen werden:

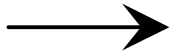
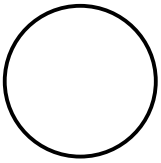
Komponente	Symbol	Bedeutung
Fluß		Beschrieben werden Bewegungen von Daten von einem Teil des Systems zu einem anderen. Die Pfeilspitze gibt dabei die Richtung des Datenflusses an, der Pfeilname die Art der Daten.
Prozeß		Ein Prozeß repräsentiert einen Teil des Systems, der Eingaben zu Ausgaben verarbeitet.
Speicher		Durch Speicher werden - im Gegensatz zu Flüssen - ruhende Daten, also z.B. Dateien, dargestellt.
Terminator		Ein Terminator steht für ein externes Objekt, mit dem das System kommuniziert.

Abb. 18 Komponenten eines Datenflußdiagrammes

Durch diese Methode können also Prozesse und Datenflüsse zwischen den Prozessen, bzw. zwischen Prozessen und externen Objekten, definiert werden; Kontrollstrukturen können jedoch nicht dargestellt werden.
Die folgende Abbildung zeigt das Gesamtsystem als DFD.

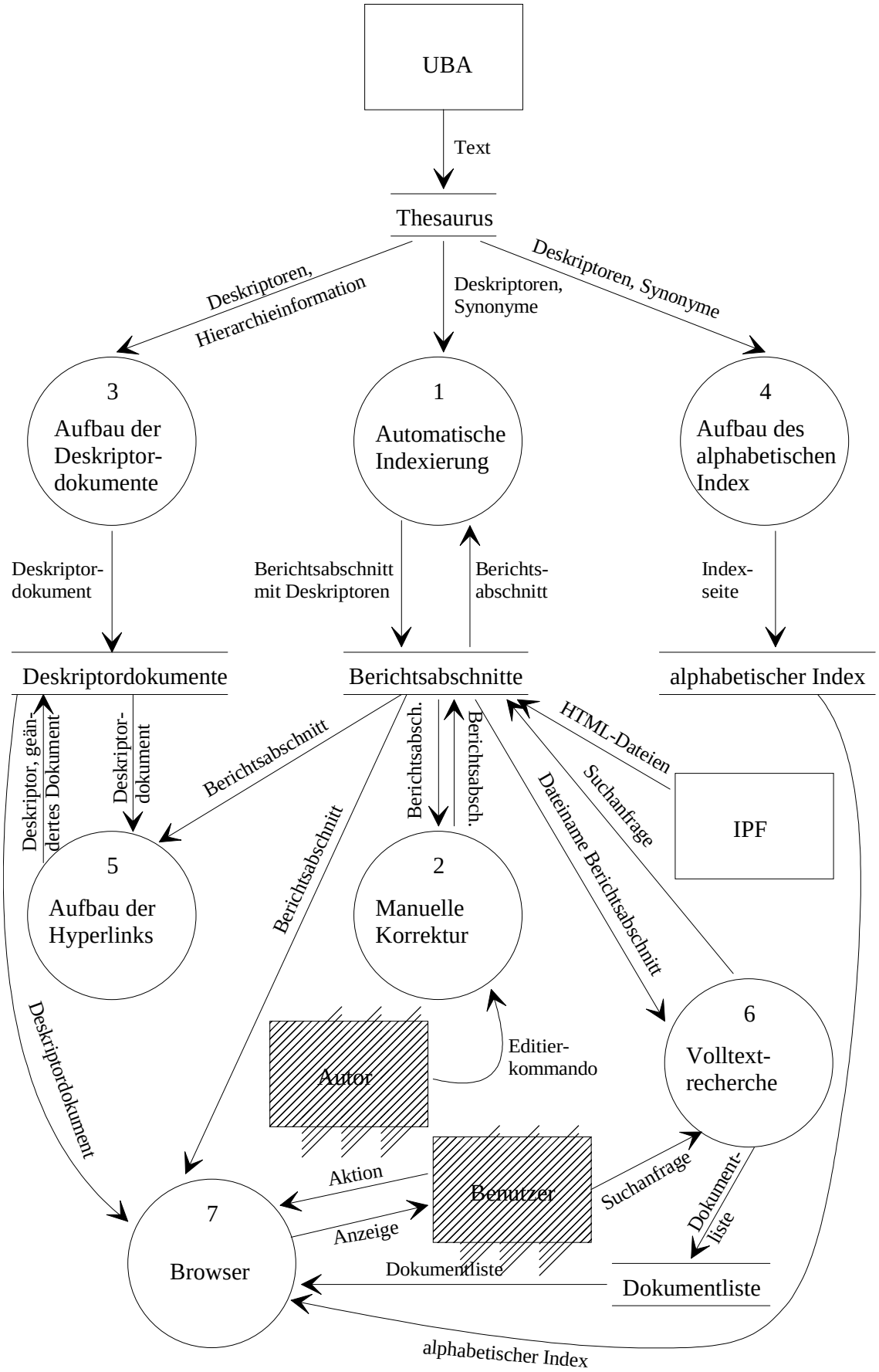


Abb. 19 Enwurf des Gesamtsystems als DFD

Die Numerierung der Prozesse gibt die Reihenfolge an, in der später die zugehörigen Programme beschrieben werden. Sie geben nicht unbedingt eine zeitliche Reihenfolge an. Zwar muß z.B. die manuelle Korrektur nach der automatischen Indexierung erfolgen, aber es ist z.B. irrelevant, ob zuerst die Deskriptordokumente oder der alphabetische Thesaurus erstellt werden.

Der in 3.4.1 beschriebene Umweltbericht findet sich wieder in den "Berichtsabschnitten". Diesen Berichtsabschnitten werden durch den Prozeß 1 "Automatische Indexierung" automatisch Deskriptoren aus dem Thesaurus (s. 3.4.2) zugeordnet.

Der Thesaurus wiederum wird dazu verwendet, um die Deskriptordokumente zu erstellen (Prozeß 3). Ein Deskriptordokument beschreibt dabei genau einen Deskriptor und seine hierarchischen Beziehungen zu anderen Deskriptoren.
Außerdem wird der Thesaurus zu einem alphabetischen Index aller Deskriptoren und Synonyme umgesetzt (Prozeß 4).
Da beim Thesaurus im Prinzip immer auf die systematische Liste zurückgegriffen wird, scheint es am übersichtlichsten, als Datenflüsse immer nur die Information anzugeben, die im jeweiligen Fall relevant ist, auch wenn der Thesaurus immer vorverarbeitet wird (und so eigentlich verschiedene "Versionen" vorhanden sind).

Die Deskriptoren in den Berichtsabschnitten können vom Autor, also demjenigen, der den Hypertext erstellt, manuell korrigiert werden (Prozeß 2).
Jedes Deskriptordokument soll Verweise auf alle Berichtsabschnitte enthalten, die mit diesem Deskriptor indexiert sind. Um diese Beziehung zwischen Berichtsabschnitten und Deskriptordokumenten herzustellen, dient der Prozeß 5 "Aufbau der Hyperlinks".
Angezeigt werden Deskriptordokumente, Berichtsabschnitte und alphabetischer Index vom "Browser" (Prozeß 7) durch Aktionen des Benutzers.
Der Benutzer kann außerdem eine Volltextrecherche starten (Prozeß 6), durch die eine Dokumentliste von Berichtsabschnitten erstellt wird, die das Suchkriterium erfüllen.

Einzelne Prozesse sollen noch weiter verfeinert dargestellt werden. Sinnvoll erscheint das bei der automatischen Indexierung und beim Aufbau der Hyperlinks. Weniger wichtig dagegen ist die weitere Verfeinerung beim Aufbau des alphabetischen Index und der Deskriptordokumente, da hier im Prinzip nur eine Datei (der Thesaurus) verarbeitet und in eine andere Form umgesetzt wird. Bei der Volltextrecherche kann weitgehend auf bereits vorhandenen Programmen aufgebaut werden. Der Prozeß "Browser" schließlich wird durch NCSA Mosaic (s. 3.3) abgedeckt; die manuelle Korrektur kann mit einem beliebigen Texteditor durchgeführt werden, schöner wäre allerdings ein spezieller Deskriptor-Editor mit Zugriff auf den Thesaurus.

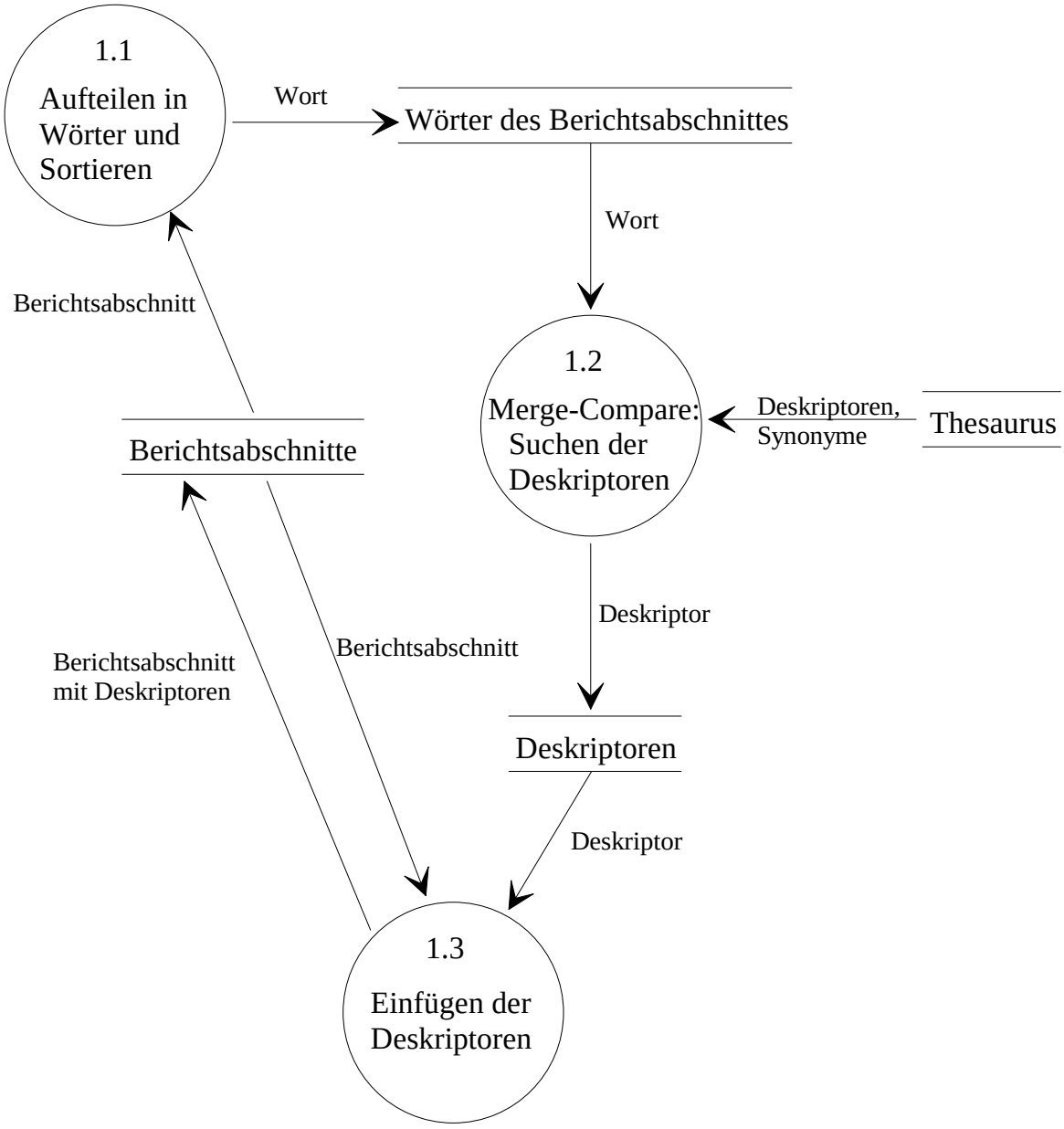


Abb. 20 Verfeinerung des Prozesses "Automatische Indexierung"

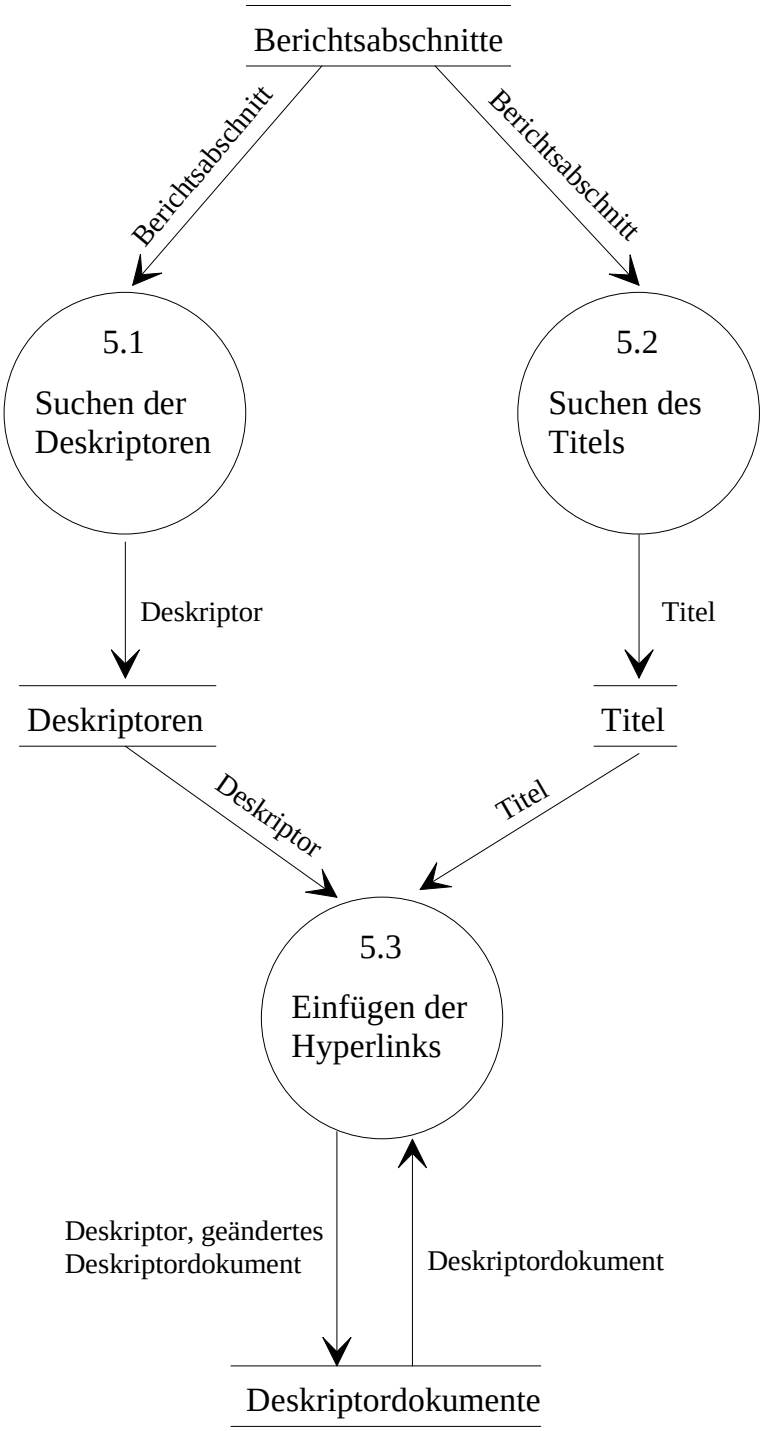


Abb. 21 Verfeinerung des Prozesses "Aufbau der Hyperlinks"

3.6 Realisierung

3.6.1 Verzeichnisstruktur für Umweltbericht und Thesaurus

Durch die Konvertierung des Umweltberichtes zu HTML-Dateien beim IPF wurden die Berichtsabschnitte bereits in eine hierarchische Struktur eingeordnet. Diese spiegelt den Kapitelaufbau des gedruckten Berichtes wider. Unterhalb des Verzeichnisses "1ud" gibt es Unterverzeichnisse für die einzelnen Kapitel, in diesen Unterverzeichnissen sind dann die einzelnen Abschnitte des jeweiligen Kapitels als HTML-Dateien enthalten.
Daneben gibt es ein Verzeichnis für die Dateien des Thesaurus ("thes"). Dieses enthält den alphabetischen Index als HTML-Datei und ein Verzeichnis für die Deskriptordokumente.
Alle Hyperlinks in den Berichtsabschnitten und Deskriptordokumenten sind relative Pfadangaben, die sich auf das übergeordnete Verzeichnis "Umweltdaten" beziehen. Dieses enthält neben den Verzeichnissen für Umweltbericht und Thesaurus auch die Startseite.

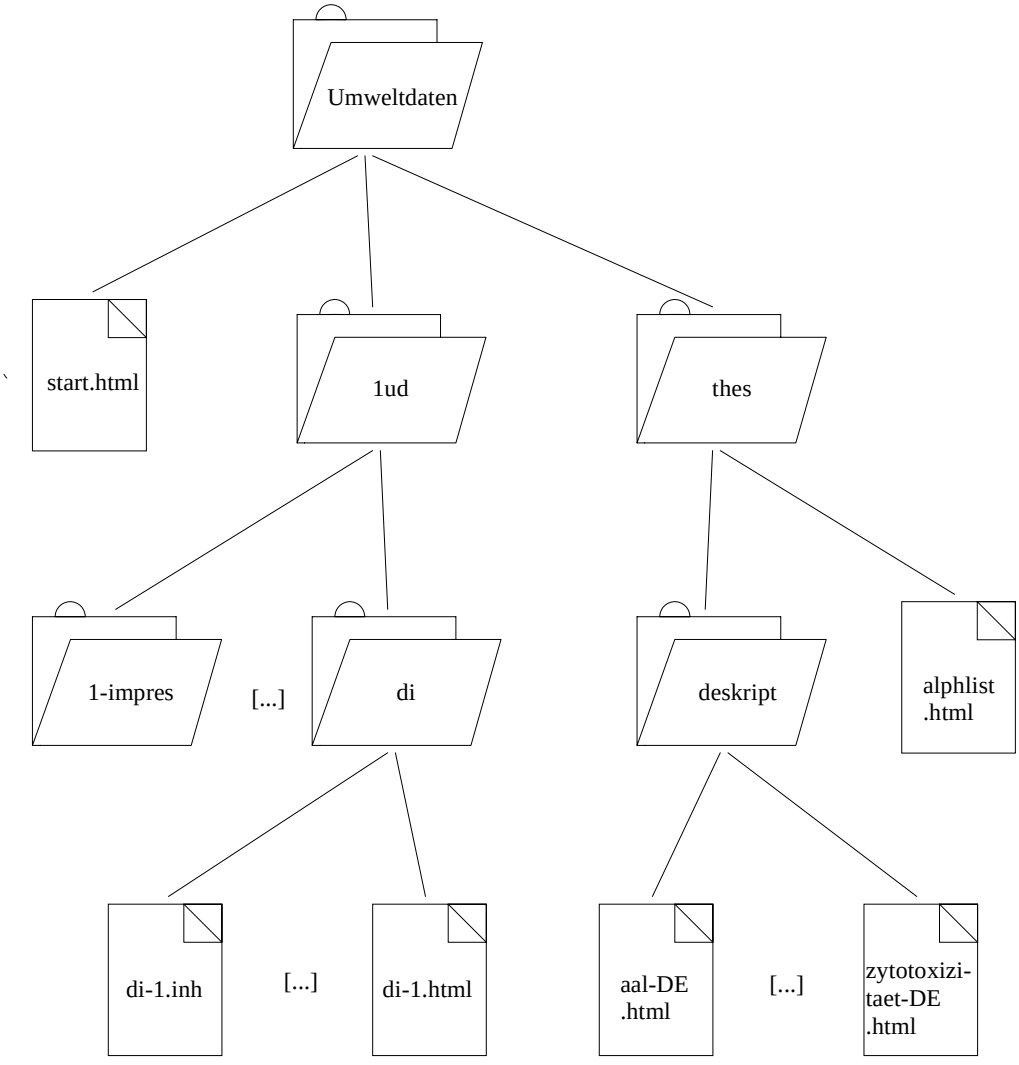


Abb. 22 Verzeichnisstruktur der Berichtsabschnitte und Deskriptordokumente

3.6.2 Automatische Indexierung der Berichtsabschnitte

Im ersten Schritt werden den bereits als HTML-Dokumente vorliegenden Berichtsabschnitten automatisch Deskriptoren aus dem Thesaurus (genauer: der systematischen Liste) zugeordnet. Dabei werden in jeden Berichtsabschnitt Zeilen eingefügt, die die Deskriptoren für diesen Berichtsabschnitt enthalten.

Alle Programme und Prozeduren zur Deskriptorvergabe werden von einem Shell-Skript ("index.csh") aufgerufen, dem als Parameter der Name der HTML-Datei übergeben wird; falls kein Parameter übergeben wird, wird der Dateiname abgefragt. Ausgabe dieses Shell-Skripts ist dann die HTML-Datei mit eingefügten Deskriptoren. Der Gesamtablauf von "index.csh" ist in Abb. 23 dargestellt.

Zuerst wird die HTML-Datei in einzelne Wörter aufgeteilt, d.h. jede Zeile der Ausgabe enthält dann ein Wort der HTML-Datei, wobei bereits eine Vorverarbeitung stattfindet. Die Aufteilung in Wörter erfolgt in mehreren Schritten mit Hilfe von Unix-Befehlen, die durch "Pipes" miteinander verbunden sind.

Im ersten Teilschritt durchlaufen die Eingabedaten ein sed-Skript ("deltag.sed"), das

- alle Zeilen bis <BODY> löscht,
- alle Hypertext-Links löscht,
- alle sonstigen HTML-Tags löscht,
- Sonderzeichen durch Blanks ersetzt und
- Umlaute durch ae (oe, ue) bzw. ß durch ss ersetzt.

Dabei werden die zu entfernenden Elemente durch reguläre Ausdrücke beschrieben.

Danach ersetzen tr-Befehle alle Blanks durch Return (d.h. jede Zeile enthält nur noch ein Wort) und alle Groß- durch Kleinbuchstaben. Die daraus entstehende Wortfolge wird mit "sort" alphabetisch sortiert und als Eingabe für ein awk-Programm ("stopwort.awk") benutzt, welches Stoppwörter, Zahlen, einbuchstabige Wörter und Leerzeilen entfernt.

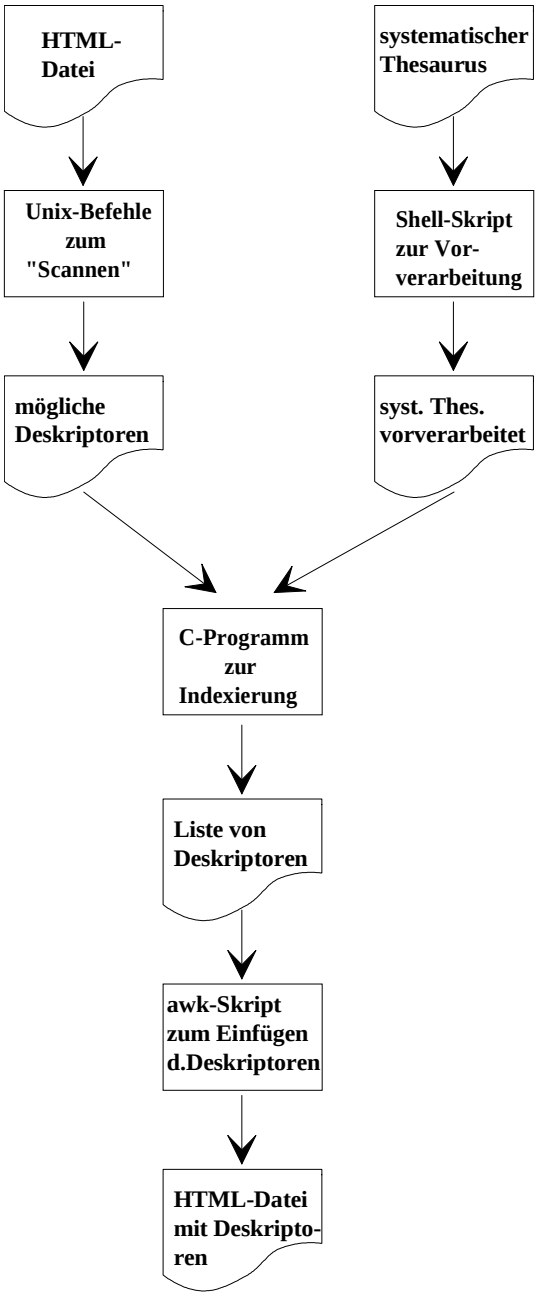


Abb. 23 Gesamtablauf der Deskriptorvergabe

Die Stoppwörter (also solche Wörter, die auf keinen Fall als Deskriptoren geeignet sind) in diesem awk-Programm werden ebenfalls durch reguläre Ausdrücke beschrieben, und zwar so, daß möglichst viele Endungen eines Wortes durch einen Ausdruck abgedeckt sind. So steht z.B. der reguläre Ausdruck *die(s(e[mnrs]?)?)?* für die Stoppwörter *die*, *dies*, *diese*, *diesem*, *diesen*, *dieser*, *dieses*. Die Stoppwörter wurden nicht automatisch erstellt, sondern mit Hilfe der Wortlisten, die durch das oben beschriebene sed- bzw. tr-Skript aus den Berichtsabschnitten angelegt wurden. Aus diesen Wortlisten wurden die häufigsten Wörter, die nicht als Deskriptoren geeignet sind, in die Stoppwortliste aufgenommen. Zwar ist die

Entfernung von Stoppwörtern nicht unbedingt notwendig, da nur solche Wörter als Deskriptoren verwendet werden, die im Thesaurus vorkommen, hat aber dennoch Vorteile. Beispielsweise wird dadurch die Liste von Wörtern, für die im Thesaurus überprüft werden muß, ob es sich um Deskriptoren handelt, erheblich kleiner (bei Tests von einigen HTML-Dateien ergab sich durchschnittlich eine Reduzierung auf ca. 30%); damit läuft die eigentliche Deskriptorvergabe dann wesentlich schneller ab.
Die Ausgabe bis zu diesem Zeitpunkt besteht also aus einer Liste von Wörtern, die möglicherweise als Deskriptoren zu verwenden sind (s. Abb 24).

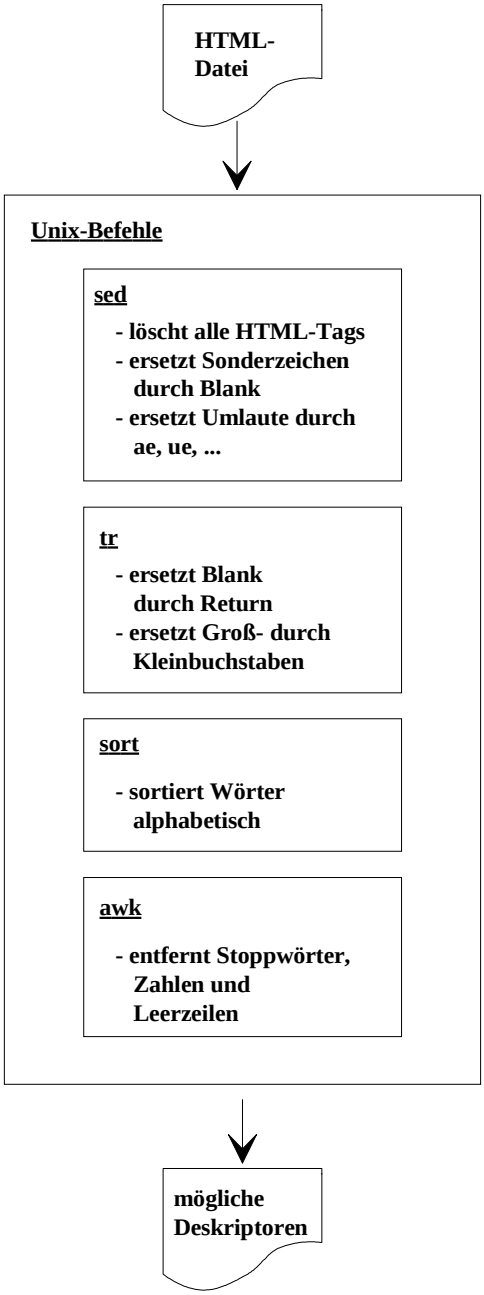


Abb. 24 Teilschritt 1 der Deskriptorvergabe

Um die Wörter dieser Liste mit den Begriffen des systematischen Thesaurus vergleichen zu können, wird dieser von einem Shell-Skript (""systkurz.csh") vorverarbeitet.
Dabei wird zum einen die Schreibweise angeglichen (Umsetzung in Kleinschreibung), außerdem werden aus dem Thesaurus die zu jedem Deskriptor aufgeführten Informationen, wie Ober- und Unterbegriffe, englische Begriffe usw., entfernt, da sie für die Deskriptorvergabe nicht relevant sind. Die zu Nicht-Deskriptoren aufgeführten und statt deren zu benutzende Synonyme und Kombinationen (s.o.) bleiben erhalten. Der vorverarbeitete systematische Thesaurus entspricht also eigentlich dem alphabetischen Thesaurus; nur sind zusätzlich abweichende Benennungen mit den dafür zu benutzenden Synonymen oder Kombinationen enthalten.

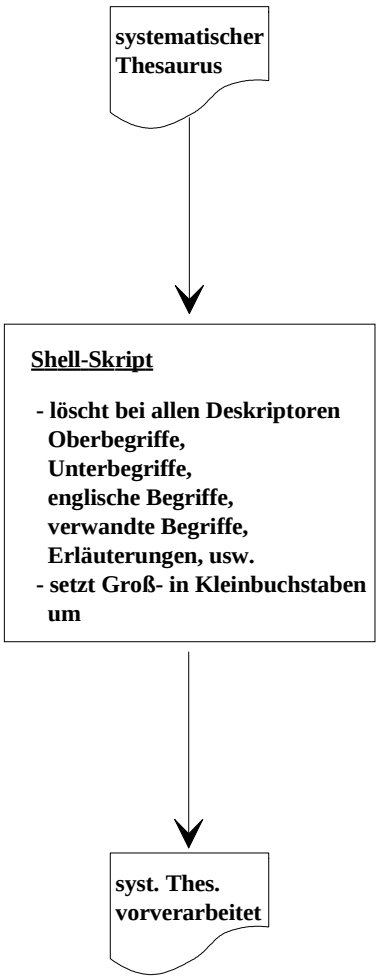


Abb. 25 Vorverarbeitung des systematischen Thesaurus

Dieser Thesaurus wird zusammen mit der Liste der möglichen Deskriptoren in einem C-Programm ("deskript.c") verarbeitet, welches aus den beiden Eingabedateien diejenigen Wörter heraussucht, die in beiden Dateien vorkommen bzw. die sich nur durch die letzten beiden Buchstaben unterscheiden (s. Abb. 26).

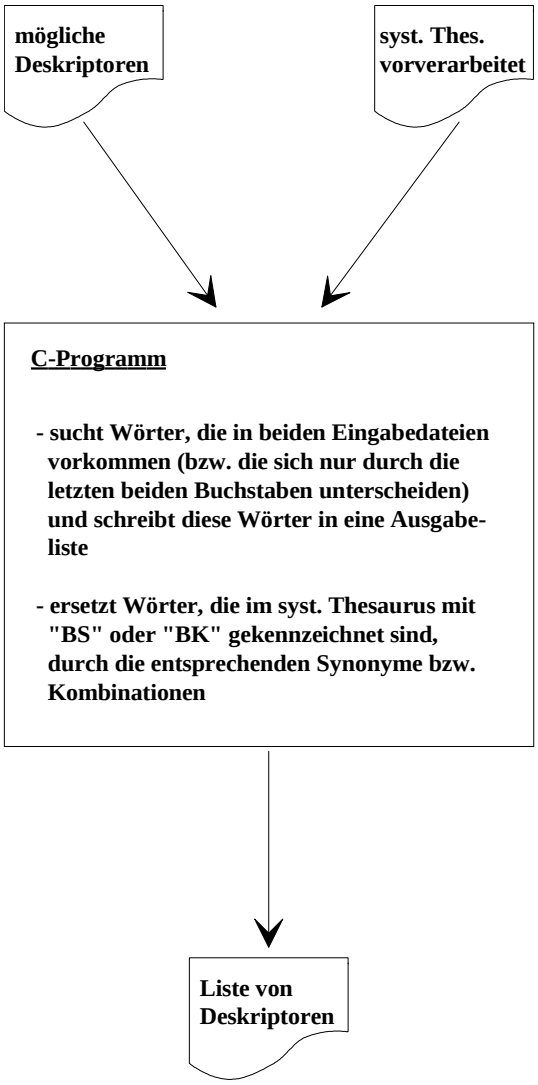


Abb. 26 Verarbeitung des Thesaurus mit den HTML-Dateien

Falls gleiche Wörter gefunden werden, die im systematischen Thesaurus mit "BS" oder mit "BK" gekennzeichnet sind, werden sie in der Ausgabe durch die entsprechenden Synonyme bzw. Kombinationen ersetzt.

Als Ergebnis liefert das C-Programm also eine Liste von Deskriptoren, wobei Deskriptoren, die im HTML-Dokument mehrfach vorkommen, auch mehrfach aufgeführt sind. Diese Liste ist wegen der durch Synonyme oder Kombinationen ersetzten Wörter nicht mehr alphabetisch sortiert.

Das Einfügen der Deskriptoren in die HTML-Datei wird von einem awk-Programm ("deskript.awk") übernommen, welchem als Parameter der Name der HTML-Datei und eine Liste der einzufügenden Deskriptoren übergeben werden. Diese Deskriptorliste baut auf der vom C-Programm ausgegebenen Liste auf; im Unterschied zu dieser ist sie jedoch nach der Häufigkeit, mit der die einzelnen Deskriptoren vorkommen, sortiert, und es gibt keine

Mehrfach-Einträge mehr. Außerdem sind Mehrwort-Deskriptoren entfernt ("einwortdeskript.awk"). Das awk-Programm fügt die Deskriptoren einmal als Links zu den Deskriptordokumenten und zusätzlich als Kommentar ein. Diese (unsichtbaren) Kommentarmarken haben den Vorteil, daß die Stelle, ab der die Deskriptoren aufgeführt sind (bzw. wo die Auflistung endet), bei späteren Verarbeitungsschritten (z.B. bei der manuellen Korrektur) eindeutig festgestellt werden kann. Beim Einfügen werden derzeit nur die sieben am häufigsten vorkommenden Deskriptoren berücksichtigt. Beim Anlegen der Hyperlinks werden die Deskriptoren als Dateinamen für die Deskriptordokumente verwendet (so verweist dann z.B. der Deskriptor "Wald" auf das Deskriptordokument "wald-DE.html").

Für die automatische Deskriptorvergabe gibt es ein weiteres Shell-Skript ("indexalle.csh"), das den oben beschriebenen Ablauf für alle Berichtsabschnitte unterhalb des Verzeichnisses "1ud" durchführt. Das Shell-Skript sucht dazu alle Berichtsabschnitte (zu erkennen an der Endung ".html") unterhalb dieses Verzeichnisses und ruft dann für jeden gefundenen Berichtsabschnitt das Shell-Skript "index.csh" auf.

Der Gesamtablauf der automatischen Indexierung soll an einem Beispiel verdeutlicht werden. Eingabedatei ist die folgende HTML-Datei, ein Berichtsabschnitt zum Thema "Ökologisches Datenbanksystem":

```
<HTML>
 <BODY>
<H1>Ökologisches Datenbanksystem </H1>
Seit 1991 läuft die Entwicklung und Erprobung eines ökotoxikologischen Datenbanksystems. Ziel ist es, über zehntausende in den letzten Jahren gewonnene Einzeldaten komplexen Auswerte- und Bewertungsverfahren zuzuführen.
<p>
Es bestehen die Möglichkeiten Abfragen räumlich, nach geographischen oder politischen Grenzen sowie projekt- oder standortbezogen durchzuführen.
<p>
Im Rahmen der Untersuchungen an den Wald - Dauerbeobachtungsflächen wurde an der Hälfte der Flächen ein Klimameßgerät mit Sensoren für Lufttemperatur und -feuchte sowie Bodentemperatur und -feuchte aufgebaut Die in einem Datenlogger aufgezeichneten Meßwerte können per RAM - Karte im Gelände ausgelesen, in die Datenbank eingelesen, ausgewertet und für die Bewertung des Witterungsverlaufs am Standort herangezogen werden.
<p>
Ein weiterer Ausbauschritt ist die Einbindung von Bewertungsalgorithmen in die Ergebnisdarstellung, so daß immer wiederkehrende gleichartige Anfragen nicht jedesmal von neuem einzeln erstellt werden müssen. Diese Bewertungsalgorithmen fußen auf einem im Verlauf des Wirkungskatasters gewonnen Regelwerkes, mit dem der Zustand eines Beobachtungsraums hinsichtlich eines oder mehrerer Parameter "benotet" werden kann. Am Ende des Integrationsprozesses der unterschiedlichen Parameter steht die Erstellung einer "Ökologischen Zustandskarte" für das Land.
<p>
<A HREF=http:/Umweltdaten/1ud/cv/cv-02_7.html> Zum folgenden Kapitel </A>
<p>
<A HREF=http:/Umweltdaten/1ud/1_titel/titel.inh> Zurück zur Eingangsseite </A>
</BODY>
 </HTML>
```

Nachdem diese Eingabedaten durch das sed-Skript, die tr-Befehle und "sort" verarbeitet wurden, ergibt sich folgende Ausgabe (wobei die einzelnen Wörter aus Platzgründen hintereinander stehen, aber eigentlich jede Zeile nur ein Wort enthält):

1991 abfragen am am an an anfragen auf aufgebaut aufgezeichneten ausbauschritt ausgelesen ausgewertet auswerte benotet beobachtungsraums bestehen bewertung bewertungsalgorithmen bewertungsalgorithmen bewertungsverfahren bodentemperatur das dass datenbank datenbanksystem datenbanksystems datenlogger dauerbeobachtungsflaechen dem den den der der der der der des des des die die die die die die die diese durchzufuehren ein ein einbindung einem einem einer eines eines eines eingelesen einzeldaten einzeln ende entwicklung ergebnisdarstellung erprobung erstellt erstellung es es feuchte feuchte flaechen fuer fuer fuer fussen gelaende geographischen gewonnen gewonnene gleichartige grenzen haelfte herangezogen hinsichtlich im im im immer in in in in integrationsprozesses ist ist jahren jedesmal kann karte klimamessgeraet koennen komplexen laeuft land letzten lufttemperatur mehrerer messwerte mit mit moeglichkeiten muessen nach neuem nicht oder oder oder oekologischen oekologisches oekotoxikologischen parameter parameter per politischen projekt raeumlich rahmen ram regelwerkes seit sensoren so sowie sowie standort standortbezogen steht ueber und und und und und unterschiedlichen untersuchungen verlauf von von wald weiterer werden werden werden wiederkehrende wirkungskatasters witterungsverlaufs wurde zehntausende ziel zustand zustandskarte zuzufuehren

Aus dieser Liste werden die Stoppwörter entfernt, übrig bleiben die folgenden Begriffe (auch hier eigentlich nur ein Begriff pro Zeile):

abfragen anfragen aufgebaut aufgezeichneten ausbauschritt ausgelesen ausgewertet auswerte benotet beobachtungsraums bestehen bewertung bewertungsalgorithmen bewertungsalgorithmen bewertungsverfahren bodentemperatur datenbank datenbanksystem datenbanksystems datenlogger dauerbeobachtungsflaechen durchzufuehren einbindung eingelesen einzeldaten einzeln ende entwicklung ergebnisdarstellung erprobung erstellt erstellung feuchte feuchte flaechen fussen gelaende geographischen gewonnen gewonnene gleichartige grenzen haelfte herangezogen hinsichtlich integrationsprozesses jahren jedesmal kann karte klimamessgeraet koennen komplexen laeuft land letzten lufttemperatur mehrerer messwerte moeglichkeiten muessen neuem oekologischen oekologisches oekotoxikologischen parameter parameter per politischen projekt raeumlich rahmen ram regelwerkes sensoren standort standortbezogen steht unterschiedlichen untersuchungen verlauf wald wiederkehrende wirkungskatasters witterungsverlaufs zehntausende ziel zustand zustandskarte zuzufuehren

Aus dieser Liste der möglichen Deskriptoren werden diejenigen Begriffe ausgegeben, die im systematischen Thesaurus als Deskriptoren vorhanden sind oder zu denen es Synonyme gibt. Im folgenden Auszug der systematischen Liste sind die für diese Beispiel relevanten Zeilen aufgelistet:

[...]
bewertungsverfahren
[...]
bodentemperatur
[...]
datenbank
[...]
datenbanksystem
  bs datenbank
[...]
dauerbeobachtungsflaeche
[...]
lufttemperatur
[...]
sensor

[...]
wald
[...]
wirkungskataster

Folgende Deskriptoren werden ausgegeben (die Wörter "Datenbanksystem" und "Datenbanksystems" aus der Liste der möglichen Deskriptoren werden durch den Deskriptor "Datenbank" ersetzt):

bewertungsverfahren
bodentemperatur
datenbank
  bs datenbank
  bs datenbank
dauerbeobachtungsflaeche
lufttemperatur
sensor
wald
wirkungskataster

Die Deskriptoren werden der Häufigkeit nach sortiert und die sieben häufigsten Deskriptoren in die HTML-Datei eingefügt, einmal als Kommentar und einmal als Hyperlink zum entsprechenden Deskriptordokument (Einfügung im Fettdruck):

<HTML>
 <BODY>
<H1>Ökologisches Datenbanksystem </H1>
<HR>
<!--DE:
datenbank wirkungskataster wald sensor lufttemperatur dauerbeobachtungsflaeche bodentemperatur
END-DE:-->
<P>Deskriptoren:
<P>
<A HREF="/Umweltdaten/thes/deskript/datenbank-DE.html">Datenbank</A>
<A HREF="/Umweltdaten/thes/deskript/wirkungskataster-DE.html">Wirkungskataster</A>
<A HREF="/Umweltdaten/thes/deskript/wald-DE.html">Wald</A>
<A HREF="/Umweltdaten/thes/deskript/sensor-DE.html">Sensor</A>
<A HREF="/Umweltdaten/thes/deskript/lufttemperatur-DE.html">Lufttemperatur</A>
<A HREF="/Umweltdaten/thes/deskript/dauerbeobachtungsflaeche-
DE.html">Dauerbeobachtungsflaeche</A>
<A HREF="/Umweltdaten/thes/deskript/bodentemperatur-DE.html">Bodentemperatur</A>
<P>
<HR>
Seit 1991 läuft die Entwicklung und Erprobung eines ökotoxikologischen
Datenbanksystems. Ziel ist es, über zehntausende in den letzten
Jahren gewonnene Einzeldaten komplexen Auswerte- und Bewertungsverfahren
zuzuführen.
[...]

Die automatisch eingefügten Deskriptoren sind manuell änderbar, d.h. es ist möglich, einen Deskriptor zusätzlich einzugeben oder einen der automatisch zugeordneten zu löschen.

3.6.3 Aufbau der Deskriptordokumente

In einem weiteren Schritt wird der Thesaurus in einen hierarchischen Hypertext aus Deskriptordokumenten konvertiert. Dabei wird aus jedem Deskriptor des systematischen Thesaurus ein Deskriptordokument erstellt, das Hyperlinks zu den Ober- und Unterbegriffen sowie zu verwandten Begriffen dieses Deskriptors enthält. Der Deskriptor selbst wird als Hyperlink zum Eintrag im alphabetischen Index angelegt. Generiert werden diese Deskriptordokumente von einem C-Programm ("desdok.c"), das dazu einen vorverarbeiteten systematischen Thesaurus benutzt. Der Ablauf ist in der folgenden Abbildung dargestellt:

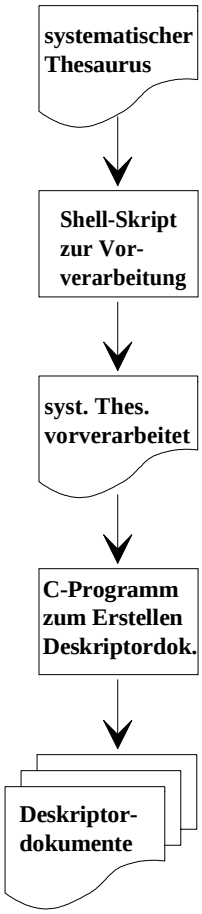


Abb. 27 Erstellen der Deskriptordokumente

Die Vorverarbeitung des Thesaurus sieht in diesem Fall so aus, daß ein awk-Programm alle Synonyme und Kombinationen sowie die englischen Übersetzungen aus dem systematischen Thesaurus entfernt; übrig bleiben also nur die Deskriptoren mit Oberbegriffen, Unterbegriffen, verwandten Begriffen usw.
Das folgende Beispiel zeigt einen Auszug aus dem Thesaurus und das Deskriptordokument, das aus diesen Zeilen erstellt wird.

```
    [...]
  Schadstoffbelastung
   OOB Chemikalien
     Umweltbelastung
   OB Schadstoff
     Umweltbelastung
   UB AOX-Wert
     Biochemischer Sauerstoffbedarf
     Chemischer Sauerstoffbedarf
     DOC
     Gesamtkohlenstoff
     Pflanzenkontamination
     Salzbelastung
     Schadstoffemission
     Schadstoffexposition
     Schadstoffgehalt
     Schadstoffimmission
     Schmutzfracht
     Schwermetallbelastung
   BF schadstoffbelastet
     Ammoniumbelastung
     Belastung (chemisch)
     Bleibelastung
     Belastung (Schadstoff)
     Cadmiumbelastung
     Grenzschadstoffbelastung
     Klaerschlammbelastung
     Kontamination
     Pestizidbelastung
     Schadstoffbelastung (hoechstzulaessig)
     Schadstoffeintrag
     SO2-Belastung
    [...]
```

wird umgesetzt zu

```
<HTML>
<HEAD>
<TITLE>Deskriptor: Schadstoffbelastung</TITLE>
</HEAD>
<BODY>
<A HREF="/Umweltdaten/thes/hthes.html"><IMG SRC="/Umweltdaten/thes/h.gif"></A>
<HR>
<H1>Deskriptor:
<A HREF="/Umweltdaten/thes/alphlist.html#schadstoffbelastung">Schadstoffbelastung</A></H1>
Dokumente:
<UL>
</UL>
<HR>
<H3>OOB:</H3>
<UL>
<LI><A HREF="/Umweltdaten/thes/deskript/chemikalien-DE.html">Chemikalien</A>
<LI><A HREF="/Umweltdaten/thes/deskript/umweltbelastung-DE.html">Umweltbelastung</A>
</UL>
<H3>OB:</H3>
<UL>
<LI><A HREF="/Umweltdaten/thes/deskript/schadstoff-DE.html">Schadstoff</A>
<LI><A HREF="/Umweltdaten/thes/deskript/umweltbelastung-DE.html">Umweltbelastung</A>
```

```
</UL>
<H3>UB:</H3>
<UL>
<LI><A HREF="/Umweltdaten/thes/deskript/aox-wert-DE.html">AOX-Wert</A>
<LI><A HREF="/Umweltdaten/thes/deskript/doc-DE.html">DOC</A>
<LI><A HREF="/Umweltdaten/thes/deskript/gesamtkohlenstoff-DE.html">Gesamtkohlenstoff</A>
<LI><A HREF="/Umweltdaten/thes/deskript/pflanzenkontamination-DE.html">Pflanzenkontamination</A>
<LI><A HREF="/Umweltdaten/thes/deskript/salzbelastung-DE.html">Salzbelastung</A>
<LI><A HREF="/Umweltdaten/thes/deskript/schadstoffemission-DE.html">Schadstoffemission</A>
<LI><A HREF="/Umweltdaten/thes/deskript/schadstoffexposition-DE.html">Schadstoffexposition</A>
<LI><A HREF="/Umweltdaten/thes/deskript/schadstoffgehalt-DE.html">Schadstoffgehalt</A>
<LI><A HREF="/Umweltdaten/thes/deskript/schadstoffimmission-DE.html">Schadstoffimmission</A>
<LI><A HREF="/Umweltdaten/thes/deskript/schmutzfracht-DE.html">Schmutzfracht</A>
<LI><A HREF="/Umweltdaten/thes/deskript/schwermetallbelastung-DE.html">Schwermetallbelastung</A>
</UL>
<H3>BF:</H3>
<UL>
<LI>schadstoffbelastet
<LI>Ammoniumbelastung
<LI>Bleibelastung
<LI>Cadmiumbelastung
<LI>Grenzschadstoffbelastung
<LI>Klaerschlammbelastung
<LI>Kontamination
<LI>Pestizidbelastung
<LI>Schadstoffeintrag
<LI>SO2-Belastung
</UL>
</BODY>
</HTML>
```

Beim Eintragen der Begriffe in die HTML-Dateien werden nur Einwortbegriffe berücksichtigt, so daß z.B. der Unterbegriff "Chemischer Sauerstoffbedarf" nicht übernommen wird. Auch bei der Deskriptorvergabe werden nur Einwortbegriffe verwendet; es kann also gar keinen Verweis auf z.B. ein Deskriptordokument "chemischer-sauerstoff-DE.html" geben.

3.6.4 Aufbau des alphabetischen Index

Bei den Deskriptordokumenten werden Hyperlinks zu einem alphabetischen Index angelegt. Dieses Verzeichnis aller Deskriptoren und Synonyme wird aus dem vorverarbeiteten systematischen Thesaurus, der auch für die Deskriptorvergabe benutzt wurde, von einem awk-Programm ("alphhtml.awk") generiert. Berücksichtigt werden bei diesem alphabetischen Index für die Deskriptoren nur Einwortbegriffe. Bei den Synonymen und Kombinationen allerdings werden auch Mehrwortbegriffe zugelassen, wenn die Begriffe, die diese Synonyme oder Kombinationen ersetzen, nur aus einem Wort bestehen. Beispielsweise wird der Eintrag "Emission (Verkehr)" übernommen, da das Synonym "Verkehrsemission" ist; der Begriff "Emission (Pro-Kopf)" wird ebenfalls in den alphabetischen Index aufgenommen, weil er

durch die Kombination "Emissionsdaten" und "Pro-Kopf-Daten" ersetzt wird. Die Auswahl der Begriffe erfolgt durch ein awk-Programm ("einwortsystkurz.awk").
Beim Generieren des alphabetischen Index wird dann jeder Deskriptor als Link zum entsprechenden Deskriptordokument angelegt. Gleichzeitig wird der Deskriptor als Zieladresse festgelegt. Dadurch erreicht man, daß man beim Springen von der Deskriptorseite zum alphabetischen Index direkt zu diesem Deskriptor gelangt und nicht erst den (sehr umfangreichen) alphabetischen Index durchblättern muß.

3.6.5 Aufbau von Hyperlinks zwischen Berichtsabschnitten und Deskriptordokumenten

Die Hyperlinks von den Berichtsabschnitten zu den Deskriptordokumenten werden bereits bei der Deskriptorvergabe angelegt, da die zu jedem Berichtsabschnitt gefundenen Deskriptoren sofort als Hyperlinks eingefügt werden. Die Verweise in der umgekehrten Richtung, also von jedem Deskriptordokument zu allen Berichtsabschnitten, die diesen Begriff als Deskriptor aufweisen, werden in einem weiteren Schritt von einem Shell-Skript ("deskriptDE.csh") eingefügt.

Der Ablauf für einen bestimmten Berichtsabschnitt sieht dabei so aus, daß zuerst alle Deskriptoren dieses Berichtsabschnittes von einem awk-Programm herausgesucht werden ("suchedeskriptor.awk"). Ein weiteres awk-Programm ("suchetitel.awk") sucht noch einen Text aus dem Berichtsabschnitt heraus, der als Hyperlink in das Deskriptordokument eingefügt werden kann. Da die zur Verfügung gestellten HTML-Berichtsabschnitte keinen Titel enthalten, wird nach der ersten Überschrift des Berichtsabschnittes gesucht. Anschließend fügt ein awk-Programm ("einfuegedeskriptor.awk") in jedes Deskriptordokument der gefundenen Deskriptoren den Dateinamen bzw. (sichtbar) den Titel des Berichtsabschnittes als Hyperlink ein.

Die einzelnen Schritte sollen noch etwas detaillierter an einem Beispiel dargestellt werden. Eingabedatei ist der Berichtsabschnitt "Ökolgisches Datenbanksystem" aus 3.6.2. Der Dateiname dieses Berichtsabschnittes wird als Parameter übergeben. Da dies der vollständige Pfadname ist, die Hyperlinks aber relativ zum Verzeichnis "Umweltdaten" eingetragen werden, wird zunächst einer Variablen ($filelink) der relative Pfadname zugewiesen, im Beispiel */Umweltdaten/1ud/cv/cv-02_6.html*.
Anschließend wird das Programm "suchedeskriptor.awk" mit diesem Berichtsabschnitt aufgerufen und erstellt folgende temporäre Datei:

datenbank
wirkungskataster
wald
sensor
lufttemperatur
dauerbeobachtungsflaeche
bodentemperatur

Das Programm "suchetitel.awk" findet als Titel des Berichtsabschnittes die Zeichenkette "Ökologisches Datenbanksystem", die der Variablen $titel zugewiesen wird.
Für jeden Deskriptor der temporären Datei wird nun das entsprechende Deskriptordokument gesucht und der Hyperlink zum Berichtsabschnitt eingefügt, wobei für die URL-Adresse die Variable $filelink und für den sichtbaren Text die Variable $titel verwendet wird. Die erste Deskriptordatei, mit der das awk-Programm "einfuegedeskriptor.awk" im Beispiel aufgerufen wird, ist also "datenbank-DE.html"; weitere Parameter sind $filelink und $titel. Nach Ablauf dieses Programmes sieht "datenbank-DE.html" folgendermaßen aus (eingefügte Zeilen im Fettdruck):

```
<HTML>
<HEAD>
<TITLE>Deskriptor: Datenbank</TITLE>
</HEAD>
<BODY>
<A HREF="/Umweltdaten/thes/hthes.html"><IMG SRC="/Umweltdaten/thes/h.gif"></A>
<HR>
<H1>Deskriptor: <A HREF="/Umweltdaten/thes/alphlist.html#datenbank">Datenbank</A></H1>
Dokumente:
<UL>
<P>
<LI><A HREF="/Umweltdaten/1ud/cv/cv-02_6.html">Ökologisches Datenbanksystem </A>
<P>
</UL>
<HR>
<H3>OOB:</H3>
<UL>
<LI><A HREF="/Umweltdaten/thes/deskript/umweltinformation-DE.html">Umweltinformation</A>
</UL>
[...]
</BODY>
</HTML>
```

Auch hier gibt es wie bei der Deskriptorvergabe ein weiteres Shell-Skript ("deskriptDEalle.csh"), das den Aufbau der Hyperlinks für jeden Berichtsabschnitt unterhalb des Verzeichnisses "1ud" durchführt.

3.6.6 Volltextrecherche

Bei der Volltextrecherche ist es möglich, die Berichtsabschnitte nach beliebigen Begriffen, also nicht nur Deskriptoren, zu durchsuchen. Für die Eingabe dieser Suchbegriffe wird ein HTML-Formular verwendet.
In dieses Formular können maximal 6 Suchbegriffe eingetragen werden; verknüpft werden können sie durch "OR" oder "AND".
Bei Anklicken von "Submit" im Formular wird das Shell-Skript "htgrep.csh" aufgerufen. Dieses Skript startet zunächst ein C-Programm zum Parsen der Variablen "QUERY_STRING", die die Eingabewerte des Formulars enthält. In Abhängigkeit von diesen Werten werden dann die entsprechenden Suchalgorithmen gestartet.

Bei der Suche im Text selbst wird auf das Unix-Programm "grep" zurückgegriffen. Bei der Verknüpfung der Suchbegriffe mit "OR" ist der Ablauf so, daß alle Berichtsabschnitte nacheinander nach den einzelnen Suchbegriffen durchsucht und die gefundenen Dateien in eine temporäre Datei geschrieben werden. Werden die Suchbegriffe mit "AND" verknüpft, werden alle Berichtsabschnitte nach dem ersten Suchbegriff durchsucht und die Dateiname in eine temporäre Datei geschrieben. Beim nächsten Suchbegriff werden nur noch die Dateien aus dieser temporären Datei durchsucht und das Ergebnis wieder in eine Datei geschrieben. So wird der Reihe nach mit allen Suchbegriffen verfahren, so daß am Ende nur noch die Dateien in der temporären Datei stehen, die wirklich alle Suchbegriffe enthalten.
Als Ausgabe, die wieder an "Mosaic" zurückgegeben wird, liefert das Skript eine HTML-Seite, in der alle Berichtsabschnitte mit den gesuchten Begriffen als Hyperlinks enthalten sind.

4. Diskussion

Darauf, was durch die Programme erreicht wurde, soll hier nicht näher eingegangen werden, da dies bereits ausführlich im letzten Kapitel erläutert wurde. Vielmehr sollen hier Probleme bei der Programmerstellung sowie mögliche Verbesserungen und Erweiterungen der bestehenden Programme angesprochen werden.

Ein erster Ansatzpunkt zu möglichen Verbesserungen ist bei den Eingabedaten selbst gegeben, sowohl beim Thesaurus als auch bei den Berichtsabschnitten.
Aus dem Thesaurus wurden für jedes Programm unterschiedliche Informationen benötigt und dann immer ein Skript vorgeschaltet, um diese Informationen herauszufiltern. Auf diese Weise existieren mehrere Dateien, die zwar alle aus dem systematischen Thesaurus abgeleitet sind, aber viel redundante Information enthalten. Hier ist zu überlegen, wie man den Thesaurus so aufbereitet, daß er für alle Programme direkt verwendet werden kann.

Bei den als HTML-Seiten aufbereiteten Berichtsabschnitten könnte schon bei der Konversion einiges verbessert bzw. noch ein Schritt vor der automatischen Indexierung zwischengeschaltet werden.
Mögliche Verbesserungen sind:

- Die HTML-Dateien sollten zumindest noch die Tags
  <HEAD>
  <TITLE></TITLE>
  </HEAD>
  enthalten. Als Titel könnte z.B. noch die erste Überschrift eingefügt werden.

- Wenn in den Texten eine zweizeilige Überschrift vorhanden war, wurde das z.B. umgesetzt zu:
  <H1>Erfassung und Überwachung </H1>
  <H1>des Zustands der Böden </H1>;
  besser wäre jedoch eine Konvertierung zu:
  <H1>Erfassung und Überwachung des Zustands der Böden</H1>,
  da Mosaic die Zeilenlänge an die aktuelle Fenstergröße anpaßt.

- Kapitel, die auch in den Inhaltsverzeichnissen der einzelnen Abschnitte noch weiter untergliedert sind (z.B. die Kapitel "Verkehr" oder "Energie" im Abschnitt "Allgemeine Daten"), sollten auch in einzelne HTML-Dateien aufgeteilt werden, da die HTML-Seite

sonst relativ lang und unübersichtlich wird; außerdem ist dadurch die Deskriptorvergabe weniger sinnvoll.

- Verweise zum Weiter- bzw. zum Zurückblättern sollten oben auf der Seite stehen, damit man nicht erst bis ans Ende der Texte gehen muß (was gerade beim schrittweisen Durchblättern sehr umständlich ist), um zum nächsten Dokument zu kommen.

- Bei den Abbildungen wäre es besser, nicht neue HTML-Seiten aufzurufen, sondern direkt z.B. zu den GIF-Dateien zu verweisen. Dann würde ein externes Programm das Bild anzeigen, und man könnte gleichzeitig den Text und die dazugehörige Abbildung sehen. Außerdem können die Hyperlinks zu den Abbildungen und Tabellen im laufenden Text an den jeweils passenden Stellen eingefügt werden, und nicht, wie bisher, erst am Ende des Textes.

- Mehr Textmarkierungen können zu HTML-Tags umgesetzt werden, z.B. Aufzählungen mit "-" oder die Auflistung der Abbildungen zu nichtnumerierten oder numerierten Listen.

Grundsätzlich könnten in den Texten selbst (zusätzlich zu den Abbildungen) noch mehr Hyperlinks enthalten sein, damit das Browsing besser unterstützt wird. Möglich wäre z.B., die Deskriptoren nicht nur in den eingefügten Zeilen als Hyperlinks anzulegen, sondern auch im Text selbst.

Verbesserungen bzw. Erweiterungen sind auch beim Algorithmus für die automatische Indexierung denkbar. Probleme gibt es noch beim Vergleich der Begriffe im Text mit den Deskriptoren und Synonymen im Thesaurus. Um hier auch andere Wortformen (z.B. Mehrzahl) in den Berichtsabschnitten zu erfassen, dürfen sich der Begriff im Berichtsabschnitt und der Begriff im Thesaurus in der Endung (genauer in den letzten beiden Buchstaben) unterscheiden. Dies ist in den meisten Fällen auch sinnvoll, ergibt aber Probleme, wenn zwei Deskriptoren im Thesaurus sehr ähnlich sind, z.B. "Wal" und "Wald". Kommt im Berichtsabschnitt z.B. der Begriff "Wald" vor, erkennt der Algorithmus schon beim Deskriptor "Wal" eine Gleichheit und ordnet diesen Deskriptor zu. Auch der Fall, daß die Mehrzahlform einen Umlaut enthält (z.B. Wald - Wälder), ist noch nicht berücksichtigt. Eventuell wäre es sinnvoll, hier auf bereits vorhandenen Algorithmen zur Stammformreduzierung aufzubauen; diese müßten dann sowohl auf den Thesaurus als auch auf die Begriffe des Berichtsabschnittes angewandt werden.

Die Stoppwortliste zum Entfernen von Begriffen, die auf keinen Fall als Deskriptoren geeignet sind, ist von Hand erstellt. Vielleicht kann diese automatisch erstellt, bzw. auf eine bereits vorhandene zurückgegriffen werden, die man dann beispielsweise in ein awk-Programm umsetzen kann. So wird z.B. bei MRESSE (1984) eine von der Zentralstelle für maschinelle Dokumentation erstellte Stoppwortliste mit ca. 2.500 Einträgen erwähnt.

Beim Einfügen der Deskriptoren in die Berichtsabschnitte werden die sieben häufigsten Deskriptoren ausgewählt; hier ist zu überlegen, ob diese Deskriptoren wirklich am besten geeignet sind. Sinnvoll könnte auch sein, alle gefundenen Deskriptoren, unabhängig von ihrer Anzahl, einzufügen; dann wäre es dem Benutzer überlassen, die weniger geeigneten Deskriptoren bei der manuellen Korrektur zu entfernen.

Die durch die automatische Indexierung vorgeschlagenen Deskriptoren können grundsätzlich manuell korrigiert werden (bevor die Hyperlinks aufgebaut werden) bzw. auch noch danach (das Programm zum Einfügen der Hyperlinks muß dann noch einmal gestartet werden). Diese Änderung geschieht aber bisher noch mit einem einfachen Texteditor. Hier könnte man sich ein weiteres Programm vorstellen, das eine komfortablere Editierung der vorgeschlagenen Deskriptoren ermöglicht. Realisiert werden kann solch ein Editor beispielsweise durch ein Tcl/Tk-Skript, das in einem Fenster den Berichtsabschnitt anzeigt, in einem weiteren die vorgeschlagenen Deskriptoren und in einem dritten Fenster alle Deskriptoren (aus dem Thesaurus), die vergeben werden können. An Funktionen müßte dieser Editor sowohl das Löschen und Einfügen von Deskriptoren ermöglichen als auch das Auswählen einer Hypertext-Datei mit einem File-Browser.
Vorstellen kann man sich auch, für alle Programme eine Menüoberfläche zu schaffen. Dort kann man dann z.B. mit einem File-Browser einen Berichtsabschnitt auswählen, für diesen die automatische Verschlagwortung starten, die Deskriptoren manuell korrigieren usw.

Weitere Verbesserungen betreffen den alphabetischen Index. Da dieser sehr umfangreich ist, dauerte es im Gegensatz zu den Berichtsabschnitten und Deskriptordokumenten relativ lange, bis er vom Browser angezeigt wird (besonders bei hoher Netzbelastung). Hier erscheint eine Aufteilung in mehrere Dateien sinnvoll, z.B. für jeden Anfangsbuchstaben eine neue HTML-Seite (mit Verweisen zum vorhergehenden und nachfolgenden Buchstaben). Dazu muß das Programm zum Erstellen des alphabetischen Index entsprechend geändert werden. Außerdem müssen die Hyperlinks zum alphabetischen Index, die beim Erstellen der Deskriptordokumente angelegt werden, so verändert werden, daß sie nicht mehr auf "alphlist.html" sondern z.B. auf "a-alphlist.html", "b-alphlist.html" usw. verweisen.

Ein weiterer interessanter Ansatzpunkt wäre es, die Deskriptordokumente erst dann zu generieren, wenn ein Deskriptor im Berichtsabschnitt angeklickt wird. Damit vom Anwählen des Deskriptors bis zum Anzeigen des Deskriptordokumentes nicht zu viel Zeit vergeht, muß ein schneller Zugriff auf den Deskriptor und seine zugehörigen Informationen möglich sein (z.B. indem der Thesaurus in einer Datenbank verfügbar gemacht wird).

Die vorgeschlagenen Änderungsmöglichkeiten können lediglich dazu beitragen, dem Endbenutzer noch mehr Navigationsmöglichkeiten, bzw. dem Autor mehr Komfort bei der Editierung zu bieten. Die prinzipielle Vorgehensweise der Konvertierung von Umweltberichten und Thesaurus wird davon jedoch nicht berührt, da der Gesamtablauf der Programme den Anforderungen durchaus gerecht wird.

5. Zusammenfassung

Die vorliegende Arbeit zeigt, wie Umweltdaten mit Hilfe von Hypertext und Methoden der automatischen Indexierung erschlossen werden können.
Konkret handelt es sich bei den Umweltdaten um einen Umweltbericht der Landesanstalt für Umweltschutz Baden-Württemberg. Dieser Bericht wird als Hypertext aufbereitet und im World Wide Web zur Verfügung gestellt. Der Zugriff auf die Informationen ist dabei nicht nur über den Text selbst gegeben, sondern auch über den Umwelt-Thesaurus des Umweltbundesamtes.

Um den Zugriff über den Thesaurus zu ermöglichen, werden den einzelnen Abschnitten des Umweltberichtes Deskriptoren aus dem Thesaurus zugeordnet. Diese Verschlagwortung erfolgt automatisch, wobei anschließend die Möglichkeit zur manuellen Korrektur der vergebenen Deskriptoren besteht. Aus dem Thesaurus selbst wird zum einen eine Liste aller Deskriptoren und Synonymbegriffe erstellt, zum anderen wird für jeden Deskriptor eine Hypertextseite erstellt, die Hyperlinks zu allen Ober- und Unterbegriffen sowie verwandten Begriffen des Deskriptors enthält. In einem weiteren Schritt wird die Verbindung zwischen diesen Deskriptordokumenten und dem Umweltbericht hergestellt, indem in jedes Deskriptordokumente Hyperlinks zu allen Abschnitten des Umweltberichtes eingefügt werden, denen der entsprechende Deskriptor zugordnet wurde.
Zusätzlich wird mit einer Volltextrecherche die Möglichkeit geschaffen, den Umweltbericht nach beliebigen Begriffen zu durchsuchen; diese Begriffe können durch eine "Und"- bzw. "Oder"-Verknüpfung kombiniert werden.
Auf diese Weise wird eine Informationsstruktur erstellt, die es ermöglicht, über verschiedene Einstiegspunkte auf die Informationen des Umweltberichtes zuzugreifen.

Abkürzungsverzeichnis

Abb.	Abbildung
ACM	**A**ssociation for **C**omputing **M**achinery
ASCII	**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange
awk	**A**ho, **W**einberger, **K**ernighan
BF	**B**enutzt **F**ür
BK	**B**enutze **K**ombination
BS	**B**enutze **S**ynonym
bzw.	beziehungsweise
ca.	circa
CERN	**C**entre **E**uropéen de **R**echerches **N**ucléaires
CGI	**C**ommon **G**ateway **I**nterface
DFD	**D**aten**F**luß**D**iagramm
d.h.	das heißt
DTD	**D**ocument **T**ype **D**efinition
EPS	**E**ncapsulated **P**ost**S**cript Format
etc.	et cetera
evtl.	eventuell
EWG	**E**uropäische **W**irtschafts**G**emeinschaft
FAW	**F**orschungsinstitut für **A**nwendungsorientierte **W**issensverarbeitung
FIZ	**F**ach**I**nformations**Z**entrum
FRESS	**F**ile **R**etrieval and **E**diting **S**y**S**tem
FTP	**F**ile **T**ransfer **P**rotocol
FZI	**F**orschungs**Z**entrum **I**nformatik
GIF	**G**raphics **I**nterchange **F**ormat
HTML	**H**yper**T**ext **M**arkup **L**anguage
HTTP	**H**yper**T**ext **T**ransfer **P**rotocol
IBM	**I**nternational **B**usiness **M**achines Corporation
IIR	**I**ntelligentes **I**nformation **R**etrieval
IKE	**I**nstitut für **K**ernenergetik und **E**nergiesysteme
IPF	**I**nstitut für **P**hotogrammetrie und **F**ernerkundung
ISO	**I**nternational **S**tandardization **O**rganization
LfU	**L**andesanstalt **f**ür **U**mweltschutz
MEMEX	**MEM**ory **EX**tender
NCSA	**N**ational **C**enter for **S**upercomputing **A**pplications
NLS	o**N**-**L**ine **S**ystem
OB	**O**ber**B**egriff

OOB	**O**berster **O**ber**B**egriff
s.	siehe
s.a.	siehe auch
sed	**s**tream **ed**itor
SGML	**S**tandard **G**eneralized **M**arkup **L**anguage
s.o.	siehe oben
sog.	sogenannte
Tcl/Tk	**T**ool **c**ommand **l**anguage/**T**ool**k**it
tr	**tr**anslate characters
u.a.	unter anderem
UB	**U**nter**B**egriff
UBA	**U**mwelt**B**undes**A**mt
UIS	**U**mwelt**I**nformations**S**ystem
URL	**U**niform **R**esource **L**ocator
usw.	und so weiter
VB	**V**erwandter **B**egriff
WAIS	**W**ide **A**rea **I**nformation **S**ervice
WMF	**W**indows **M**eta**F**ile
WWW	**W**orld **W**ide **W**eb
z.B.	zum Beispiel

Literaturverzeichnis

BERNERS-LEE, T., R. CAILLIAU :
World-Wide Web.
Vorabdruck eines Beitrages zur Konferenz "Computing in High Energy Physics 92",
Annecy, September 1992

BERNSTEIN, M. :
An apprentice that discovers hypertext links,
in "Hypertext: Concepts, Systems and Applications".
eds. A. RIZK, , N. STREITZ, J. ANDRÉ,
Cambridge University Press, Cambridge 1990

BUSH, V. :
As We May Think.
Atlantic Monthly 176:101, 1945

COVE, J.F., B.C. WALSH :
Online text retrieval.
Information Processing and Management 24:31, 1988

DECEMBER, J., N. RANDALL :
The World Wide Web Unleashed.
Sams Publishing, Indianapolis 1994

GILSTER, P. :
Der Internet-Navigator.
Carl Hanser Verlag, München, Wien 1994

GRAU, O. :
Alles integriert: Informationssurfen im World Wide Web.
c't 6:76, 1994

KNORZ, G. :
Automatische Generierung inferentieller Links in und zwischen Hyperdokumenten,
in "Experimentelles und praktisches Information Retrieval".
Hrsg. R. KUHLEN,
Universitätsverlag Konstanz, Konstanz 1992

KRAUSE, J. :
Intelligentes Information Retrieval - Rückblick, Bestandsaufnahme und Realisierungschancen,
in "Experimentelles und praktisches Information Retrieval".
Hrsg. R. KUHLEN,
Universitätsverlag Konstanz, Konstanz 1992

KUHLEN, R. :
Hypertext: ein nicht-lineares Medium zwischen Buch und Wissensbank.

Springer-Verlag, Berlin, Heidelberg 1991

KUHLEN, R. :
Hypertext und Information Retrieval - mehr als Browsing und Suche,
in "Experimentelles und praktisches Information Retrieval".
Hrsg. R. KUHLEN,
Universitätsverlag Konstanz, Konstanz 1992

LÜCK, W., W. RITTBERGER, M. SCHWANTNER :
Der Einsatz des Automatischen Indexierungs- und Retrievalsystems (AIR) im Fachinformationszentrum Karlsruhe,
in "Experimentelles und praktisches Information Retrieval".
Hrsg. R. KUHLEN,
Universitätsverlag Konstanz, Konstanz 1992

LUSTIG, G., H. ZIMMERMANN :
Indexierung,
in "Lexikon der Informatik und Datenverarbeitung".
Hrsg. H.-J. SCHNEIDER,
Oldenbourg Verlag, München 1991

MAIER, G., A. WILDBERGER :
In 6 Sekunden um die Welt: Kommunikation über das Internet.
Addison-Wesley, Bonn [u.a.] 1993

MRESSE, M. :
Information Retrieval - Eine Einführung.
B. G. Teubner, Stuttgart 1984

NIELSEN, J. :
Hypertext and Hypermedia.
Academic Press, San Diego 1990

RADA, R. :
Hypertext: from text to expertext.
McGraw-Hill, Berkshire 1991

REARICK, T.C. :
Automating the Conversion of Text Into Hypertext,
in "Hypertext/Hypermedia Handbook".
eds. E. BERK, J. DEVLIN,
McGraw-Hill, New York 1991

RINER, R. :
Automated Conversion,
in "Hypertext/Hypermedia Handbook".
eds. E. BERK, J. DEVLIN,
McGraw-Hill, New York 1991

SALTON, G., M.J. MCGILL :

Information Retrieval - Grundlegendes für Informationswissenschaftler.
McGraw-Hill, Hamburg, New York [u.a.] 1987

SCHNUPP, P. :
Hypertext.
Handbuch der Informatik,
Oldenbourg Verlag, München 1992

UMWELTBUNDESAMT (Hrsg):
Umwelt-Thesaurus.
Informationsschrift zum Umwelt-Thesaurus, Stand: 14.10.1993

WILSON, E. :
Links and structures in hypertext databases for law,
in "Hypertext: Concepts, Systems and Applications".
eds. A. RIZK, , N. STREITZ, J. ANDRÉ,
Cambridge University Press, Cambridge 1990

YOURDON, E. :
Moderne strukturierte Analyse.
Wolfram's Fachverlag, Attenkirchen 1992